THE UNIVERSITY
of EDINBURGH

A CMOS System for High Throughput Fluorescence Lifetime Sensing using Time Correlated Single Photon Counting

David Tyndall

Doctor of Philosophy

The University of Edinburgh

October 2013

ABSTRACT

Fluorescence lifetime sensing using time correlated single photon counting (TCSPC) is a key analytical tool for molecular and cell biology research, medical diagnosis and pharmacological development. However, commercially available TCSPC equipment is bulky, expensive and power hungry, typically requiring iterative software post-processing to calculate the fluorescence lifetime. Furthermore, the technique is restrictively slow due to a low photon throughput limit which is necessary to avoid distortions caused by TCSPC *pile-up*.

An investigation into CMOS compatible multimodule architectures to miniaturise the standard TCSPC set up, allow an increase in photon throughput by overcoming the TCSPC *pile-up* limit, and provide fluorescence lifetime calculations in *real-time* is presented. The investigation verifies the operation of the architectures and leads to the selection of optimal parameters for the number of detectors and timing channels required to overcome the TCSPC pile-up limit by at least an order of magnitude.

The parameters are used to implement a low power miniaturised sensor in a 130 nm CMOS process, combining single photon detection, multiple channel timing and embedded pre-processing of the fluorescence lifetime, all within a silicon area of < 2 mm^2. Single photon detection is achieved using an array of single photon avalanche diodes (SPADs) arranged in a digital silicon photomultiplier (SiPM) architecture with a 10 % fill-factor and a compressed 250 ps output pulse, which provides a photon throughput of > 700 MHz. An array of time-interleaved time-to-digital converters (TI-TDCs) with ≈ 50 ps resolution and no processing dead-time records up to eight photon events during each excitation period, significantly reducing the effect of TCSPC pile-up. The TCSPC data is then processed using an embedded centre-of-mass method (CMM) pre-calculation to produce single exponential fluorescence lifetime estimations in *real-time*.

The combination of high photon throughput and *real-time* calculation enables advances in applications such as fluorescence lifetime imaging microscopy (FLIM) and time domain fluorescence *lifetime* activated cell sorting. To demonstrate this, the device is validated in practical bulk sample fluorescence lifetime, FLIM and simulated flow based experiments. Photon throughputs in excess of the excitation frequency are demonstrated for a range of organic and inorganic fluorophores for minimal error in lifetime calculation by CMM (< 5 %).

SUPERVISION & EXAMINATION

Supervised by

Dr. Robert Henderson
and
Dr. David Renshaw

School of Engineering
The University of Edinburgh
Kings Buildings
Mayfield Road
Edinburgh
EH9 3JL

Examined by

Dr. Alessandro Esposito

MRC Cancer Cell Unit
The Hutchison/MRC Research Centre
The University of Cambridge
Hills Road
Cambridge
CB2 0XZ

and

Dr. Alister Hamilton

School of Engineering
The University of Edinburgh
Kings Buildings
Mayfield Road
Edinburgh
EH9 3JL

on

Thursday 29th August 2013

DECLARATION OF ORIGINALITY

I hereby declare that the research recorded in this thesis (excluding the exceptions stated below) and the thesis itself originated with and was composed entirely by myself.

The design of the CMOS sensor as outlined in Chapter 4, *High Throughput Fluorescence Lifetime Sensor*, was assisted by Dr. Bruce Rae, at the time of The University of Edinburgh.

The fluorescence lifetime experimental results presented in Sections 5.7 and 5.8 were captured with the assistance of Dr. Jochen Arlt, Abigail Johnston and Katy Dickinson of the University of Edinburgh.

Other people who have contributed to this work are acknowledged and/or referenced appropriately within.

This work has not been submitted for any other degree or professional qualification except as specified.

David Tyndall

ACKNOWLEDGEMENTS

I have had the pleasure of working with many highly skilled, professional and knowledgable people throughout the duration of this research, without whom the creation of the many revisions of the *SiPM_CMM* device and its supporting development platform, as well as the practical demonstration of the system, would not have been possible.

I would first like to thank my academic supervisor at the University of Edinburgh, Dr. Robert Henderson, whose ideas, guidance and dedication has been invaluable throughout this research. Special thanks must also go to Dr. David Renshaw for providing me with valuable advice at the most crucial moments. To my colleagues in the CMOS Sensors and Systems group – Drs. Bruce Rae, Justin Richardson, Richard Walker, Day-Uei (David) Li and Nikola Krstajic – thank you for your practical contributions, our many knowledgable discussions, and most importantly your enthusiasm for the research that absorbed all of our lives. I must also thank Dr. Jochen Arlt, who not only assisted with laboratory access, but gave up his time and gave a considerable effort to help successfully demonstrate the functionality of the system.

The work presented in this thesis would not have been possible without funding from the U.K. Engineering and Physical Sciences Research Council (EPSRC). Furthermore, I am thankful for the strong relationship established between our research group and STMicroelectronics – and in particular Lindsay Grant, Dr. Andrew (Drew) Holmes and Steve East – who provided access to and support of their 130 nm CMOS imaging process for the development and fabrication of the many devices. I must also thank the Royal Society of Edinburgh for awarding me a grant to carry out research at the Université Joseph Fourier in Grenoble, France. The warm welcome I received and the hospitality shown by Dr. Antoine Delon and Meike and Janek Landsberg made my three months in Grenoble thoroughly enjoyable; *merci beaucoup* and *vielen Dank!*

There are many other people and groups who have made a noteworthy contribution to this research at some stage over the last four years: the members of the MEGAFRAME consortium, in particular Drs. David Stoppa and Edoardo Charbon and their research groups from Fondazione Bruno Kessler (FBK), Trento, Italy and École Polytechnique Fédérale de Lausanne (EPFL), Switzerland; Dr. Simon Ameer-Beg and his research group at Kings College London (KCL); Katy Dickinson and Abigail Johnston (Physics and Astronomy), Susan Kivlin and Peter Lomax (Engineering) and Dr. Nhan Pham (Biological Sciences), all at the University of Edinburgh;

my new employer Dialog Semiconductor, and in particular John Armitage, for demonstrating flexibility and understanding in allowing me to complete this work; and finally, all of the authors and co-authors of the many publications [1–11] that I have been involved with.

My final words of gratitude must go to my friends and family; I look forward to spending more time with all of you as this period of my life draws to a close. In particular I would like to thank my Mum for keeping me grounded and helping me keep perspective at all times; my Dad and brother Chris for teaching me to spend the time and effort to approach things the *right* way; and my brother Niall for his assistance with all things mathematical, I wish you every success in your own doctoral research endeavours. However, my biggest thanks of all must go to the person who has undoubtedly endured the most over these four long years – Lisa, thank you for all of the selfless sacrifices you have made and for your kindness, belief, patience and dedicated support. You are the most special person in the world to me and I will forever be in your debt.

ACRONYMS

ADC	Analogue-to-Digital Converter
API	Application Programming Interface
APD	Avalanche Photodiode
CMM	Centre of Mass Method
CLCC	Ceramic Lead-less Chip Carrier
CMOS	Complimentary Metal Oxide Semiconductor
CSOM	Confocal Scanning Optical Microscopy
CFD	Constant Fraction Discriminator
CW	Continuous Wave (laser)
DCR	Dark Count Rate
DNA	Deoxyribonucleic Acid
DNL	Differential Nonlinearity
DOE	Diffractive Optical Element
DAC	Digital-to-Analogue Converter
FPGA	Field Programmable Gate Array
FSM	Finite State Machine
FIFO	First-In First-Out
FACS	Fluorescence-Activated Cell Sorting
FCS	Fluorescence Correlation Spectroscopy
FLIM	Fluorescence Lifetime Imaging Microscopy
FRET	Förster Resonance Energy Transfer
FWHM	Full Width at Half Maximum
fNIRS	functional Near Infrared Spectroscopy
GRO	Gated Ring Oscillator
GUI	Graphical User Interface
GFP	Green Fluorescent Protein
HTS	High-Throughput Screening
IRF	Instrument Response Function
INL	Integral Nonlinearity
IEM	Integration for (lifetime) Extraction Method

IP	Intellectual Property
LSB	Least Significant Bit
LFSR	Linear Feedback Shift Register
LUT	Lookup Table
MLE	Maximum-Likelihood Estimation
MCP	Micro-Channel Plate
MSB	Most Significant Bit
NIM	Nuclear Instrument Module
PLL	Phase-Locked Loop
PMT	Photomultiplier Tube
PDE/PDP	Photon Detection Efficiency / Probability
PET	Positron Emission Tomography
PCB	Printed Circuit Board
PVT	Process, Voltage and Temperature
RNG	(MATLAB) Random Number Generator
RLD	Rapid Lifetime Determination
RNA	Ribonucleic Acid
SOM	Scanning Optical Microscopy
STI	Shallow Trench Isolation
SiPM	Silicon Photomultiplier
SPAD	Single Photon Avalanche Diode
SPC	Single Photon Counting
SLM	Spatial Light Modulator
SoC	System on Chip
TCSPC	Time Correlated Single Photon Counting
TI	Time-Interleaved
ToF	Time of Flight
TTTR	Time-Tagged Time-Resolved
TAC	Time-to-Analogue Converter
TDC	Time-to-Digital Converter
TTL	Transistor-Transistor Logic
USB	Universal Serial Bus
VDL	Vernier Delay Line

PREFACE

The research documented in this thesis largely took place during the latter half of a three year period of practical work between 2009 and 2012. Many other achievements have been made during this time period that are not included in the following six chapters and these are briefly described below.

The original topic of research was on parallel fluorescence correlation spectroscopy (FCS) using a state of the art 32×32 array of CMOS single photon avalanche diodes (SPADs) [12] developed as part of the E.U. funded MEGAFRAME project. This research began by creating a custom application platform for the device that could be used for experimentation with limited technical engineering support. The development platform consists of a printed circuit board (PCB), field programmable gate array (FPGA), firmware and a large software application that was developed from the ground up. The platform was tailored to not only support FCS but also time-correlated single photon counting (TCSPC), thanks to each SPAD in the device being paired with its own *embedded* time to digital converter (TDC).

A collaborative effort between myself and the group of Dr. Antoine Delon at the Université Joseph Fourier in Grenoble, France, began to demonstrate the system for parallel FCS data acquisition. Despite successfully capturing experimental results of bulk samples and cells, it transpired that a collaboration between University College Los Angeles (UCLA) and Politecnico di Milano was working on *exactly* the same topic and had begun publishing heavily [13] just as we were submitting our own articles. At this point, we published and presented the novel work from our research [4, 5] and I began the process of exploring ideas for a new research topic using the knowledge and experience gained from developing this powerful scientific tool.

During the time spent developing the platform, I was involved with the *tape-out* of an updated 32×32 sensor with a unified TDC architecture[1]. I was also responsible for supporting the platform – and its newly fabricated sensor – as a fluorescence lifetime sensing system using TCSPC. This work involved integration of the centre-of-mass method (CMM) algorithm developed by Dr. Day-Uei (David) Li onto FPGA; and collaboration with the group of Dr. Simon Ameer-Beg at Kings College London, where experiments were carried out to

[1] The original device contained two 16×32 sub-arrays with different TDCs to test alternative approaches.

seed a funding application to the U.K. Biotechnology and Biological Sciences Research Council (BBSRC). This £750,000 grant application was accepted and work began in 2011 on improving the platform to perform multiplexed multiphoton fluorescence lifetime microscopy for real-time imaging of protein-protein interactions by Förster resonance energy transfer (FRET). Furthermore, due to growing interest from other parties wishing to use the system, an internal technology exploitation grant was applied for and subsequently awarded to build, market, sell and support the platforms as a cutting edge research tool.

Thanks to the all of the work described above, my new topic of research began to materialise. It became clear to me that the technology used for these solid state multi-pixel TCSPC arrays could be used to advance the state of the art in *single* channel TCSPC instrumentation for high throughput fluorescence lifetime sensing experiments. This culminated in the design of six chip variants over the course of 18 months, the most recent of which is described in great detail throughout this thesis.

CONTENTS

FIGURES

TABLES

1 INTRODUCTION

1.1 Improving Time Domain Fluorescence Lifetime Sensing using Advanced CMOS Technologies

Fluorescence lifetime sensing is a key analytical tool for many applications throughout the life sciences [14]. As well as being independent of probe concentration, illumination intensity and emission wavelength, it can be used to acquire more quantitative information about physiological parameters such as pH levels [15] as well as O_2 [16] and Ca^{2+} concentrations [17]. Time-correlated single photon counting (TCSPC) is the most precise technique for measuring fluorescence lifetime in the time domain [18] and is performed by repeatedly timing fluorescence emission from a sample with respect to a synchronised pulsed optical excitation source to build a histogram of the lifetime decay. However, a major limitation of this approach is the restrictively low photon throughput limit of $\approx$ 1 – 10 % of the excitation rate [19], which is necessary in order to avoid distortion of the decay histogram due to various forms of photon event losses, commonly referred to as TCSPC *pile-up* [20]. Furthermore, typical TCSPC apparatus consists of many discrete components – including a PC for data processing – resulting in a bulky, expensive, complex and power hungry system.

Recent developments in single-photon detection and integrated timing on standard complimentary metal-oxide-semiconductor (CMOS) processes for high-resolution, solid-state Time-of-Flight (ToF) image sensors, is enabling important advances in cell-biology research, medical diagnosis and pharmacological development [21]. The most powerful of these devices is a 160$\times$128 array, where each *pixel* contains a detector and timing electronics, providing over 20,000 parallel TCSPC channels in an area less than 1.5 cm^2 [22]. However, this device produces data at over 25 Gb/s, creating a major data-handling bottleneck which further reduces the photon throughput. Furthermore, it has a low fill factor of $<$ 2 %, making detection inefficient without a complex optical set up. However by using the technology developed for these sensors, it is possible to: miniaturise standard single-channel TCSPC apparatus; develop a new integrated System on Chip (SoC) architecture to provide a method to overcome the TCSPC *pile-up* limit; and to exploit the powerful signal processing capability of advanced CMOS to reduce data-rates by performing the fluorescence lifetime calculation on-chip.

1.2 Background

1.2.1 Overview

This introductory chapter begins by providing background information to form a solid understanding of the underlying theory and core concepts that appear throughout this thesis. The aim here is also to provide a clear motivation for the work described within and is achieved by examining and answering the following fundamental questions:

- *What* is fluorescence lifetime?
- *How* is fluorescence lifetime measured?
- *Why* is fluorescence lifetime an important analytical tool?
- *What* are the limitations of fluorescence lifetime measurement techniques?

1.2.2 Fluorescence Lifetime

Fluorescence is the emission of light (or other electromagnetic radiation) by a molecule following the absorption of light (or other electromagnetic radiation) by that molecule [14]. Therefore to fluoresce, a molecule must first be stimulated by external radiation. This stimulation is commonly performed using a light source, such as a laser, that is focused onto the sample under observation. The absorption of this incident radiation, in the form of a photon, excites the molecule from its equilibrium ground-state to a vibrational energy level in an excited singlet state (S_{1-n}). Two examples of this are shown in the Jablonski diagram in Figure 1.1, where the dashed blue arrow shows absorption to the excited singlet state S_1, and the solid blue arrow shows absorption to a higher excited singlet state S_n.

The molecule then relaxes through a number of vibrational energy levels and/or other excited singlet states until it is at the lowest energy level in S_1. This occurs via two processes: vibrational relaxation to reach the lowest vibrational energy level within the current state (purple arrows), and internal conversion to transfer between excited singlet states (orange arrow). Both of these processes consist of non-radiative energy level transitions. A short time after excitation the molecule will return to its ground-state (S_0), releasing energy in the form of a photon. It is this final radiative energy level transition that gives rise to fluorescence, as shown by the green arrow in Figure 1.1. Radiative and non-radiative energy level transitions are denoted by straight and sinuous arrows, respectively, in the figure.

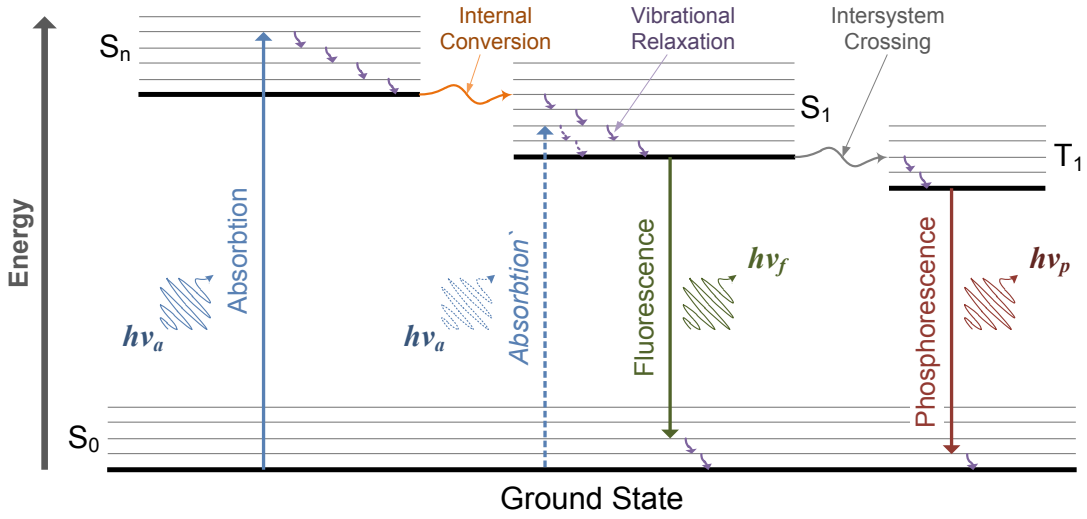


Figure 1.1: *Jablonski Diagram showing absorption, fluorescence and phosphorescence as well as internal conversion, vibrational relaxation and intersystem crossing. [14]*

It is the time delay between the absorption of radiation and the emission of a photon that leads to the concept of fluorescence lifetime. Assuming an ideal, infinitely short (δ-function) excitation pulse, this time delay follows a transient exponential decay distribution, given by Equation 1.1, where I_0 is the intensity at time $t = 0$ and τ is the lifetime of the sample. The fluorescence lifetime can then be defined quantitatively as the average time a molecule spends in its excited state(s) after the absorption of a photon, including the time taken for all non-radiative processes to complete, before emitting a secondary photon.

$$I(t) = I_0 \cdot e^{-t/\tau} \tag{1.1}$$

The losses incurred during the vibrational relaxation and internal conversion processes normally cause the emitted photon to be of lower energy – and hence longer in wavelength – than the absorbed photon. This phenomenon is known as the Stokes shift and can be used to separate excitation and fluorescence emission using optical filters. Other molecular relaxation mechanisms exist, which involve the transition of the molecule to one or more triplet excited states (T_{1-n}) before returning to S_0, either directly (phosphorescence – shown on the right of Figure 1.1) or via an excited singlet state (delayed fluorescence – not shown). The transition from singlet to triplet states is referred to as intersystem crossing and is a non-radiative energy level transition.

1.2.3 Measurement Techniques

It is possible to determine fluorescence lifetime either directly in the time domain, or indirectly in the frequency domain. The latter is performed by measuring amplitude and/or phase changes between a sinusoidal optical excitation and the detected sinusoidal emission. This is shown conceptually in Figure 1.2a, where the blue wave represents the excitation and the green wave represents the detected emission. The phase change (ϕ) and modulation factor change ($m = \frac{B/A}{b/a}$) are related to the fluorescence lifetime as shown by Equations 1.2 and 1.3, respectively, where ω is the angular frequency of the excitation and emission [14]. If the fluorescence lifetime contains only a single exponential decay characteristic, the result of both phase and amplitude modulation will be the same. However, for multi-exponential decay characteristics, measurements must be taken over a number frequencies and the results fit to a set of dispersion relationships to determine all lifetime coefficients [23].

$$tan(\phi) = \omega \cdot \tau \tag{1.2}$$

$$m = \frac{B/A}{b/a} = \left(1 + \omega^2\tau^2\right)^{-\frac{1}{2}} \tag{1.3}$$

There are two predominantly used time domain techniques for fluorescence lifetime sensing: time-gating and time-correlated single photon counting (TCSPC). Time-gated fluorescence lifetime detection performs single photon counting (SPC) during two or more time windows that are synchronised with a pulsed excitation source [24]. The windows are typically nanoseconds wide and can be overlapping or non-overlapping. An example of time-gating is shown in Figure 1.2b for two non-overlapping gates, represented by the two grey boxes labeled **A** and **B**. A variation of the rapid lifetime determination (RLD) calculation technique [25] is typically used on captured time-gated data to produce an estimation of the fluorescence lifetime. The TCSPC technique on the other hand uses a picosecond resolution timer, akin to a stopwatch, which accurately measures the time difference between a pulsed excitation source and detection of the emitted photon [18, 19]. In most implementations, TCSPC provides better timing resolution than time-gating, however only one measurement is possible per excitation period, which is the primary cause of the TCSPC *pile-up* limit [19] that will be discussed in detail in Section 2.3. By repeating this measurement many times, it is possible to build a histogram of the fluorescence decay, as shown by the vertical green lines in Figure

1.2b. The resulting histogram is fitted to variations of Equation 1.1 using iterative non-linear least-squares method or maximum-likelihood estimation algorithms [26, 27]. Time-gated detection with RLD can be performed much faster than TCSPC with these iterative non-linear algorithms, but at the expense of reduced timing resolution. These and other fluorescence lifetime calculation techniques will be discussed in more detail in Section 2.6.

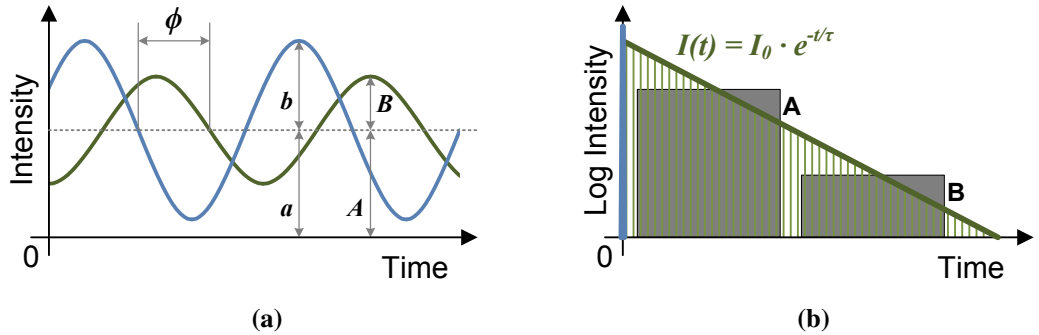


Figure 1.2: *Measuring fluorescence lifetime* **(a)** *in the frequency domain and* **(b)** *in the time domain using time-gating and TCSPC.*

In the early days of practical fluorescence lifetime sensing, the frequency domain was the preferred choice due to limitations with the speed of pulsed light source technologies, which exposed the pile-up limitation in TCSPC due to its relationship with the excitation rate. With the advent of sub-nanosecond pulsed light sources, the popularity of time domain analysis has increased significantly. Often, this is due to the fact that the sample under observation is exposed to less excitation light compared to the *always-on* light sources used in the frequency domain, which can result in sample damage caused by photo-bleaching effects. Furthermore, it has been shown that the time domain offers significantly better separability of multiple lifetime targets than the frequency domain for fluorescence lifetime tomography [28]. For these reasons, this work will focus solely on time domain fluorescence lifetime sensing approaches. Nevertheless, frequency domain techniques are still capable of higher overall photon throughput [29–32], which is necessary for some applications that will be introduced in Section 1.2.5.

System miniaturisation has been achieved in CMOS using time-gated fluorescence lifetime sensing – with one [33], two [34] and four [35] parallel time-gated counters – aimed at using RLD to reduce data bandwidth and processing requirements. However such approaches are less photon-efficient than TCSPC, due to the trade-off required between the number of time-gated counters available in parallel and the duration that each is enabled. For high accuracy, many

small time-gates are required at the expense of either a large hardware cost (parallel acquisition) or inefficient data collection (serial acquisition). Conversely, for high efficiency a small number of wide time-gates are required at the expense of accurate timing resolution (as used for two-gate RLD). Due to the inefficiencies with time-gating, this thesis will focus solely on developing an integrated hardware approach to high-throughput fluorescence lifetime sensing using time-correlated single photon counting (TCSPC), which is the most common time domain technique available, providing 100 % photon efficiency if operated below the pile-up limit, unlike gated techniques [18]. TCSPC forms the core theme of Chapter 2, which will describe its theory and limitations in more detail before introducing techniques to overcome those limitations using advanced CMOS technologies.

1.2.4 Fluorescent Probes

The majority of molecules do not naturally fluoresce, so it is therefore necessary to tag the molecules or sample of interest with a fluorescently labeled probe (or probes) of high chemical specificity, making it possible to observe the environmental or molecular properties of the sample [36]. Since the first practical demonstration of fluorescence lifetime sensing, probes have been developed to quantitatively measure a variety of different physiological environmental parameters such as pH levels, oxygen (O_2) concentrations and calcium ion (Ca^{2+}) concentrations. Examples of probes used to measure these specific parameters are outlined together with characteristic lifetimes in Table 1.1.

Parameter	**Probe**	**Typical lifetime values**				**Ref(s)**
pH level	SNAFL-1	2.7 ns	@ pH 7.9,	3.4 ns	@ pH 5.8	[15, 37]
	Acridine	14 ns	@ pH 7.9,	26.3 ns	@ pH 5.8	[37]
O_2 concentration	RTDP	775 ns	(free),	425 ns	@ 300 µM	[16, 38]
Ca^{2+} concentration	Quin-2	1.3 ns	(free),	11.6 ns	(bound)	[17, 39]
	Indo-1	0.3 ns	(free),	1.7 ns	(bound)	[39]
	Fura-2	1.72 ns	(free),	2.1 ns	(bound)	[39]

Table 1.1: *Examples of environmentally sensitive fluorescence lifetime probes.*

As can be seen in the table, typical lifetime values lie in the low nanosecond region (SNAFL-1, Quin-2, Indo-1 and Fura-2), however can extend towards the microsecond region, as is the case for RTDP. The use of RTDP to perform O_2 concentration sensing highlights one of the problems caused by the TCSPC *pile-up* limit, which is due to the extended excitation period required to allow the decay to be resolved with minimal error. Assuming an excitation period

requirement of $T > 7\tau$ to provide a fluorescence lifetime calculation error of less than 0.5 % [10], Indo-1, Quin-2, Acridine and RTDP can be resolved with excitation source repetition rates running at up to ≈ 80 MHz, 12 MHz, 5 MHz and 180 kHz, respectively. Further assuming an optimistic pile-up limited photon-throughput of 10 % of the excitation frequency [19] and ignoring any brightness limitations, the maximum photon throughputs of these fluorophores will then range from 8 MHz for Indo-1 down to only 18 kHz for RTDP. Additionally, when two or more fluorophores are measured simultaneously, the excitation source must be slow enough for the longest lifetime to be resolved and so all fluorophores must have photon rates equal to or below that of the acceptable TCSPC pile-up limit for the required excitation repetition rate.

The measurement of multiple fluorophore lifetimes simultaneously using TCSPC is further complicated when the maximum brightness levels of each fluorophore is taken into consideration. Table 1.2 shows examples of the maximum photon-rate per molecule achievable from a set of fluorophores together with the excitation power required to reach this peak [40]. Consider the case where lifetimes of GFP, Rhodamine 6G and Cy3 are to be measured simultaneously with a 10 MHz repetition rate excitation source and an optimistic pile-up limited photon-throughput of 10 % of the excitation frequency [19]. Further assuming a linear relationship between the maximum available photon rate per molecule and the excitation power, the column on the right of the table shows the power required to keep the photon-rate below the pile-up limit. As can be seen, Rhodamine 6G would require a reduced excitation power of 12.2 mW, which would have the effect of reducing the available photon-rates of the GFP sample to 256 kHz and the Cy3 sample to 172 kHz, which is significantly below their optimum. This can be viewed as a reduction of dynamic range and creates an inefficiency in data acquisition which leads to reduced certainty in the results due to the lack of photons, or more likely an increase in acquisition time by a factor of almost six in this example.

Fluorophore	**Max. Photon-rate per molecule [Hz]**	**Excitation Power [mW]**	***Excitation Power to produce a 1MHz photon-rate [mW]***[1]
Fura-2	1.25×10^3	0.2	n/a
Fluorescein	1.50×10^5	5	n/a
GFP	2.00×10^6	94	47
Rhodamine 6G	5.50×10^6	67	12.2
Cy3	2.40×10^7	1,700	70.8

Table 1.2: *Examples of Fluorescent molecules maximum photon throughput. [40]*

[1] Assuming a linear relationship between photon-rate and excitation power.

1.2.5 Applications

As previously introduced, fluorescence lifetime sensing is a key analytical tool for many important applications in the life sciences, particularly in molecular and cell biology research, medical diagnosis and pharmacological development [21]. This is in part due to its independence from probe concentration, illumination intensity and emission wavelength [14], but more importantly it provides improved knowledge of processes at the molecular level in comparison to standard fluorescence intensity. This thesis will focus primarily on two fluorescence lifetime sensing configurations – fluorescence lifetime imaging (FLIM) and flow based cytometry or sorting – each of which is capable of performing a range of important applications. However, both are affected by the TCSPC pile-up limit, particularly in the case of the latter. This section will introduce the two techniques and briefly discuss the effect that the TCSPC pile-up limit has on each.

Fluorescence Lifetime Imaging Microscopy (FLIM)

Fluorescence lifetime imaging microscopy (FLIM) using TCSPC is typically performed by scanning optical microscopy (SOM), using a single laser and detector together with a 2-dimensional scanning system. The scanning system allows the laser to be moved across the sample under observation whilst providing positional (X-Y) information to produce an image. This can be performed using standard wide-field microscopy, confocal scanning optical microscopy (CSOM) for enhanced spatial resolution [41] or two-photon microscopy for enhanced depth penetration enabling 3D sectioning [42]. FLIM can also be performed using solid-state imaging arrays in time-gated [43] and frequency domain [29, 44] configurations, each of which has its own disadvantages as discussed in Section 1.2.3. Furthermore, scanned TCSPC FLIM provides the most complete and detailed data for post-experiment analysis.

One common application of FLIM is to localise environmental changes in pH levels and ion concentrations, using the probes introduced in Section 1.2.4 and detailed in Table 1.1. However, a more challenging application of FLIM is the localisation of protein-protein interactions using fluorescence lifetime to measure Förster (or fluorescence) Resonance Energy Transfer (FRET), the use of which has grown substantially in the past decade [30, 36, 45, 46]. FRET occurs when two fluorophores (a donor and an acceptor) of high quantum yield and substantial spectral overlap have a proximity in the low nanometer scale. Energy transfer between the donor and acceptor causes fluorescence of the donor to be quenched and the acceptor to be increased,

resulting in a drop in lifetime of the donor. By tagging proteins, which interact at the same distances, with specific donor-acceptor pairs, these interactions can be localised using FLIM. All of these measurement methods (pH level, ion concentration, FRET, etc.) can be used in a variety of disciplines and application areas such as pharmacological drug development using high-throughput screening (HTS) [47, 48], analysis of DNA microarrays [11, 49], cell based medical diagnostics [44] and many more specialised applications [46].

The results from a FLIM-FRET experiment of EGFP-Cy3 labeled silica beads, captured using two-photon scanned TCSPC, is shown in Figure 1.3 [30]. The second image highlights the improved contrast and detail available using FLIM over the intensity image (left), whilst a two-exponential fitting algorithm allows the computation of the relative lifetime fraction to determine FRET efficiency (right). However – in addition to the pile-up limit – due to the nature of acquiring an image sequentially pixel-by-pixel, scanned TCSPC FLIM is a relatively slow process and can be up to three orders of magnitude slower than equivalent time-gated or frequency domain acquisition [31]. In the cases where faster acquisition speeds are required, a spatially parallel imaging modality – such as time-gating or frequency domain – is typically used. Data acquisition is further hampered when bright or long lifetime fluorophores are measured simultaneously with dim or short lifetime fluorophores, as the experimental set up must be tailored to the worst case (as described in Section 1.2.4). Moreover, performing iterative non-linear analysis on such a large data set is not trivial and can take many minutes or hours depending on the complexity of the fitting equation and the resolution of the image [50]. However, if it is possible to overcome the TCSPC pile-up limit, enhance photon-throughput and remove the requirement for post-processing data, the efficiency of data collection would improve and hence image acquisition times for TCSPC FLIM would reduce.

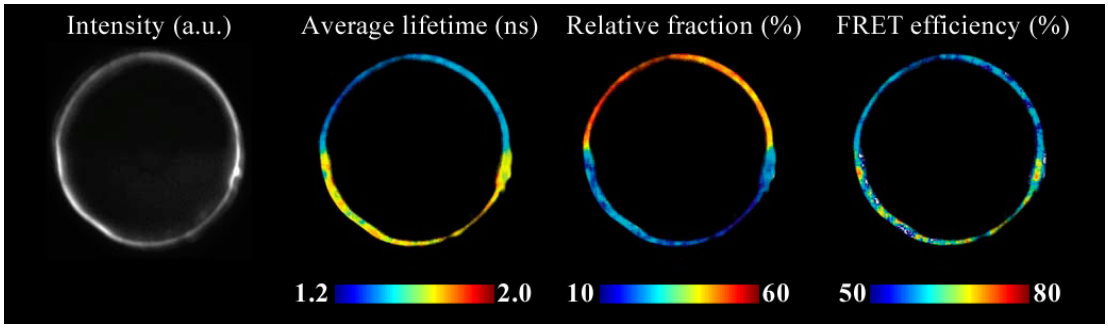


Figure 1.3: *Two-photon scanned TCSPC FLIM-FRET experiment of EGFP-Cy3 labeled silica beads. [30]*

Cytometry and Sorting

Lifetime can also be used as an analysis parameter in fluorescence flow cytometry [51] or fluorescence-activated cell sorting (FACS) [52] either independently or in addition to the traditional techniques of capturing intensity, spectral and/or scattering information. However, flow based fluorescence lifetime sensing is a demanding application and requires a high photon throughput due to the limited time each individual particle is available for analysis. This is typically below one millisecond, caused by flow rates in excess of thousands of particles per second in modern apparatus. Furthermore, FACS has the additional requirement of low-latency fluorescence lifetime calculation in order to perform the sorting function in *real-time*. Moreover, the same issue is applicable here as with FLIM, where the frequency and power of the excitation source must be tailored to the worst-case fluorophore in any given experiment (see Section 1.2.4). Due to these demands on throughput and *real-time* calculation, frequency domain is the technique of choice for these applications [53] and although TCSPC can be used for flow analysis [54], to the best of our knowledge it has not been used for sorting.

As well as being suitable for improving the classification and characterisation of a parameter set for a sample of cells, fluorescence lifetime flow cytometry, and FACS in particular, are used to perform high throughput screening (HTS) on bead based assays involving antibodies, enzymes and toxins for pharmacological drug development [55] as well as DNA and RNA analysis of cells for clinical cancer diagnosis [56]. Performing HTS with this technique has the advantage of automated sorting, which is not possible with FLIM where an additional step is required to extract the *hits*. This extraction requires complex, manually operated equipment and is a very time consuming task, completing only 50 to 100 *hit* retrievals per day [57].

Neither flow cytometry nor FACS require a microscope and so the fluorescence lifetime system is ideally suited to miniaturisation. Scaling down the hardware requirements has been developed and implemented in the time domain both as a time-gated microsystem with integrated micro-LED excitation source [34] and as a TCSPC fluorometer with an integrated laser diode excitation [58]. However, neither of these approaches would be capable of performing fluorescence lifetime based flow cytometry or FACS at a fast enough rate due to the throughput limitations caused by pile-up and the latency in calculating a fluorescence lifetime result. Therefore time domain fluorescence lifetime based flow cytometry and FACS is only achievable by overcoming the pile-up limit to increase the available photon throughput and producing a *real-time* fluorescence lifetime calculation for sorting decisions.

1.3 Research Aims

The primary aims of this research are threefold; firstly, to miniaturise the standard TCSPC set up, to not only reduce costs, but to simplify the laboratory equipment required to perform fluorescence lifetime experiments. A typical TCSPC experimental set up is shown in Figure 1.4, where the detector, timing, processing and delay-line are all to be integrated into a single CMOS sensor with a low silicon area (< 2 mm^2) and power consumption (< 10mW). Secondly, to develop a new integrated System on Chip (SoC) architecture to overcome the single channel TCSPC pile-up limit by at least an order of magnitude, to provide photon throughputs in excess of the excitation frequency. In doing so, this will provide an alternative to frequency domain techniques for applications where high-throughput is required, enabling a reduction in acquisition time for FLIM and the ability to capture sufficient photons per particle in flow cytometry. Finally, to exploit the powerful signal processing capability of an advanced CMOS technology to perform a fluorescence lifetime calculation on chip to reduce data-rates, providing *real-time* information to enable sorting to be performed in time domain flow cytometry and allowing the post-processing requirements of FLIM to be significantly reduced.

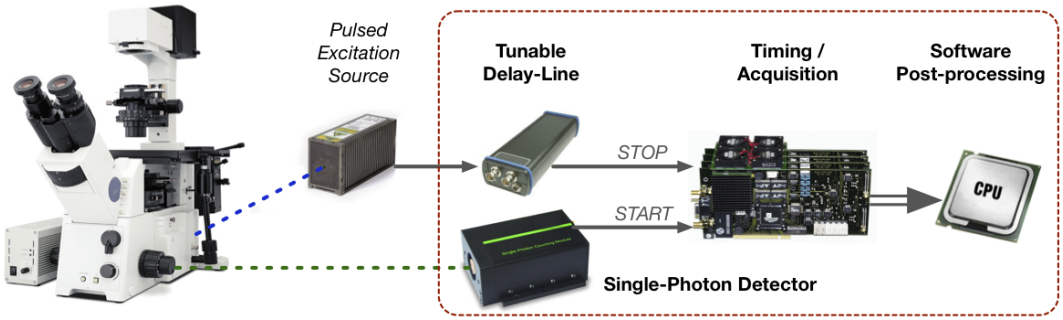


Figure 1.4: *A typical fluorescence lifetime TCSPC set up showing the components to be integrated into a single CMOS sensor within the dashed box.*

In order to reach these goals, an in-depth study is necessary into TCSPC, the pile-up limit and techniques to overcome it. It is also necessary to model the chosen technique to confirm its suitability before fabrication of any CMOS device. Furthermore, test and characterisation of the completed sensor is necessary to validate the theory, modelling and design before the results of any practical laboratory based experiments can be accepted. More specifically, the aims of the research are:

- The exploration of integrated, CMOS compatible architectures to miniaturise the standard TCSPC set up, overcome the pile-up limit and perform processing on-chip.
- The development of a modelling environment to study the chosen architecture and make informed decisions on the design variables within it.
- The design of an integrated sensor built on the modelled architecture to be fabricated in an advanced CMOS process.
- The design of a development platform to test and characterise the sensor, to include:
  - a Field Programmable Gate Array (FPGA) and USB based PCB.
  - FPGA firmware to interface with the sensor.
  - a software application with graphical user interface (GUI) to control the sensor as well as capture, process and visualise data from it.
- The validation of the sensor in the form of test and characterisation.
- The demonstration of the system as a replacement for the standard TCSPC set up in practical fluorescence lifetime applications requiring increased throughput.

1.4 Enabling Factors

The research into high performance single photon avalanche diode (SPAD) structures [59] and picosecond resolution, low-latency, time-to-digital converters (TDCs) [12] in standard CMOS processes for both fluorescence lifetime and Time-of-Flight (ToF) ranging applications has enabled this research to focus solely on developing and demonstrating an advanced sensor architecture to improve the efficiency of TCSPC. Furthermore, the parallel development of fluorescence lifetime calculation algorithms [9] provided flexibility in making the correct choice for the implementation of an embedded processing circuit.

Thanks to the overlapping research interests between our CMOS Sensors and Systems Group and the Collaborative Optical Spectroscopy Micromanipulation & Imaging Centre (COSMIC) at the University of Edinburgh, invaluable time and resource was made available in their microscopy laboratory. Finally, without the collaborative agreement between The University of Edinburgh and the imaging division of STMicroelectronics, access to an advanced 130 nm imaging specific CMOS design kit and fabrication facility would not have been possible. Many individuals are responsible for the above contributions, for which the reader is referred to the acknowledgements section at the beginning of the thesis.

1.5 Contribution to Knowledge

To the best of our knowledge, this thesis describes the first implementation of a low-power miniaturised TCSPC sensor, integrating single-photon sensitive detection, picosecond resolution timing, embedded data processing and synchronisation correction on a single CMOS substrate measuring just 1.5×1.3 mm. Furthermore, it is the first CMOS implementation of a centre-of-mass method (CMM) fluorescence lifetime estimation pre-calculation, which provides single exponential lifetime calculations in *real-time* with negligible processing latency. We also believe that the resulting sensor is the first single component hardware solution to overcoming the severe TCSPC pile-up limit by over an order of magnitude, allowing photon throughputs in excess of the excitation frequency without the requirement for additional hardware or complex software based post-processing algorithms. The combination of embedded *real-time* calculation and higher photon throughput enables advances in a number of application areas such as FLIM and time domain fluorescence *lifetime* activated cell sorting. The increased throughput is validated in practical bulk sample fluorescence lifetime, FLIM and simulated flow based experiments. In these experiments, a photon throughput of up to ten times the excitation frequency is demonstrated for a ≈ 16 ns lifetime fluorophore and minimal error in lifetime calculation by CMM (≈ 5 %).

Even without the time-correlated and embedded lifetime calculation circuitry, the single photon sensitive detector is capable of a photon throughput of up to ≈ 700 MHz, thanks to a 250 ps output pulse-width, enabling advances in high-dynamic range (uncorrelated) single-photon counting (SPC) for applications such as scanned optical microscopy (SOM) fluorescence intensity imaging. The sensor itself has been presented at a world-leading solid-state circuits conference [3], from which it was invited to a special edition of a respected biomedical circuits journal [2]. Furthermore, a journal article describing the modelling of the sensor architecture has recently been published [1]. Additional publications have also come directly or indirectly from the work carried out during this research, both as the primary author [4] and as a co-author [6–11]. The publications directly relating to this work are available in full in Appendix B. It is our firm belief that more exciting and novel work is still to come with the use of the sensor in cutting edge research for medical diagnosis and pharmacological applications, promising many more publications at leading conferences and in high-impact journals.

1.6 Thesis

The thesis aims to demonstrate how integrating single photon detection, picosecond precision timing and embedded processing on a single CMOS substrate can overcome the TCSPC pile-up limit and increase throughput for fluorescence lifetime sensing. This is achieved by investigating the current state of the art TCSPC techniques to overcome the pile-up limit, followed by system modelling to validate and direct the design variables of an architecture for a custom CMOS sensor. Block level implementation details of the sensor are then provided together with circuit simulation results to further validate the chosen architecture and design choices.

Experimental results are demonstrated using the fabricated device to back up the modelled and simulated expectations, with particular focus given to the improved performance and efficiency possible at high photon rates. These results aim to prove the suitability of the device as a sensor to be used in cutting edge biomedical research for medical diagnostics and pharmacological development. Finally, the work is concluded by outlining possible future directions and improvements of the device. A summary outline of each chapter in the remainder of this thesis is given below.

Chapter 2: Time-Correlated Single Photon Counting

This chapter presents a review of state of the art TCSPC techniques and architectures, including hardware methods to overcome the pile-up limit. A review of the technologies available in deep sub-micron CMOS to implement these architectures is also included, before the chapter concludes by describing the selection of an architecture to overcome the pile-up limit.

Chapter 3: Pile-up in an Integrated TCSPC Architecture

This chapter begins by describing an approach to modelling the chosen architecture from Chapter 2 using a MATLAB environment. The model is then used to investigate each variable in the design before the chapter concludes by making proposals for the implementation of the device in silicon.

Chapter 4: High-Throughput Fluorescence Lifetime Sensor

This chapter describes the system and block level design of the custom CMOS sensor. To validate the correct operation of the design, results of circuit simulations are presented. Furthermore, the additional hardware requirements necessary to enable suitable test and characterisation of the device are described.

Chapter 5: Sensor Test and Characterisation

This chapter begins with a description of the development platform required to bring up the fabricated device before characterisation results of the many aspects of the sensor are presented. Finally, the results from practical laboratory based microscopy experiments are shown, proving the ability of the device to overcome the TCSPC pile-up limit by over an order of magnitude.

Chapter 6: Conclusions

The final chapter presents a summary and critical discussion of the work performed and results produced during this research. Finally, additional applications that are suited to the device are presented before concepts for improvements in the sensor architecture are described.

2
TIME-CORRELATED SINGLE PHOTON COUNTING

2.1 Introduction

Chapter 1 has given a brief introduction to the theory of fluorescence lifetime – including a description of different measurement techniques and an overview of some important applications that it is used for. This chapter will focus solely on the state of the art hardware instrumentation and technologies required to perform fluorescence lifetime sensing using time-correlated single photon counting (TCSPC) – concentrating on techniques to allow it to achieve higher photon throughputs than are currently possible, with a view to integrating a custom architecture into a miniaturised CMOS sensor. It begins by presenting a review of the standard single-channel TCSPC set up, including a discussion of the distinct hardware components required for synchronisation, detection, timing and data handling. The single major limitation of TCSPC – the pile-up limit, caused by the inability of the hardware to process every photon event – is then described before techniques sometimes used to alleviate the problem are introduced.

The chapter continues by looking at different hardware architectures of what are commonly referred to as multi-module systems – where different configurations of multiple detectors and/or timers are integrated together – discussing the advantages and disadvantages of each with a focus on the ability to overcome the pile-up limit. The different technologies available to implement these architectures in CMOS – single-photon detection using single-photon avalanche diodes (SPADs) and picosecond timing using gated ring oscillator time-to-digital converters (GRO-TDCs) – are then described. Following this, the different techniques available to perform a *real-time* fluorescence lifetime calculation are investigated to decide on the most appropriate for integration as an embedded digital processing circuit. Finally, the chapter is concluded with proposals for the general architecture configuration, technology and calculation technique to be modelled before being implemented in an advanced CMOS process.

2.2 Single Channel TCSPC

2.2.1 Overview

Time-correlated single photon counting (TCSPC) is an experimental technique used to measure – with picosecond accuracy – the time between a pulsed optical excitation and a returning photon. It can be used for Time-of-Flight (ToF) ranging – which is similar to radar but using visible or infrared light in place of radio waves – where the returning photon is caused by the reflection of the excitation light from an object under observation. The distance can then be calculated using the speed of light, $c \approx 3 \times 10^8$ ms^{-1}. However, TCSPC is primarily used for fluorescence lifetime sensing [19, 60, 61], where the returning photon is emitted from a molecule soon after after the absorption of a photon from the excitation pulse, as described in Section 1.2.2. It is the most precise technique used for measuring fluorescence lifetime, being 100 % photon efficient when operated within the pile-up limit and providing the best time resolution [18]. The remainder of this chapter will focus on its use for this sole application. However, it should be noted that different applications, such as ranging, suffer from the same drawbacks of the TCSPC technique and approaches introduced in this chapter to overcome these limitations may be suitable in a number of applications other than fluorescence lifetime.

The TCSPC instrumentation set up for fluorescence lifetime consists of a number of different discrete hardware components (modules), as shown in a typical configuration in Figure 2.1. The major functions of interest to this work – synchronisation, detection, time digitisation and data-handling – are highlighted in grey, red, green and blue, respectively, and will be discussed in detail in the following sections.

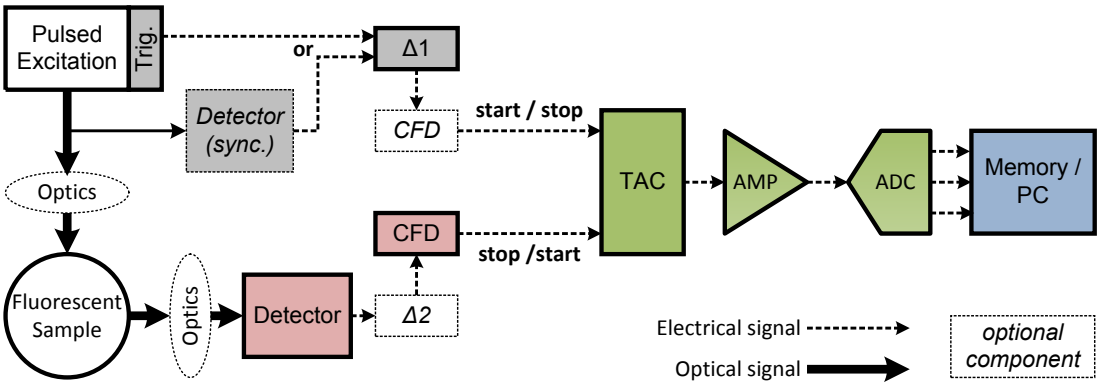


Figure 2.1: *Typical TCSPC set-up highlighting synchronisation (grey), detection (red), timing (green) and data-handling (blue).*

2.2.2 Synchronisation

To perform a time measurement, in addition to the asynchronous electrical output from the single photon detector in the fluorescence path, an electrical signal must also be provided that is synchronous with the optical excitation [18]. This excitation synchronisation is achieved in one of two ways: either a small portion of the excitation light is re-directed towards a secondary detector which provides a sufficiently fast (low jitter) electrical output pulse; or the excitation source itself provides its internal synchronisation signal – running at the same rate as the optical pulses (shown as 'Trig.' in Figure 2.1) – as an output. An additional component is then typically used to delay this excitation synchronisation pulse to compensate for any optical and/or electrical offsets in the system and to allow the fluorescence decay to be positioned as required within the range of the timer. This is shown by the component labelled $\Delta 1$ in Figure 2.1. A delay can also be placed in the detection path, as shown by $\Delta 2$, but this is less common.

As a consequence of only recording, on average, up to one photon arrival event in ten excitation cycles to remain within the pile-up limit, TCSPC time converters are typically operated in a *reverse(d) start-stop* mode, where the incoming photon event starts the timer and the excitation synchronisation stops it [19] . This flexibility is highlighted in Figure 2.1, where the excitation and detection paths are shown as *start* or *stop* signals. The *reverse* mode then has the advantage of keeping power consumption low, but more importantly it ensures that the analogue to digital converter (ADC) is less likely to be in reset when a photon event does arrive. It is also advantageous to configure the delay ($\Delta 1$) so that the timer is stopped by the synchronisation pulse corresponding to the optical excitation that caused the detected emitted photon. This improves system timing performance by removing jitter from the excitation source. The technique is shown in the timing diagram in Figure 2.2. The reverse start-stop mode means that the resulting histogram is also reversed and so must be taken into consideration for data analysis.

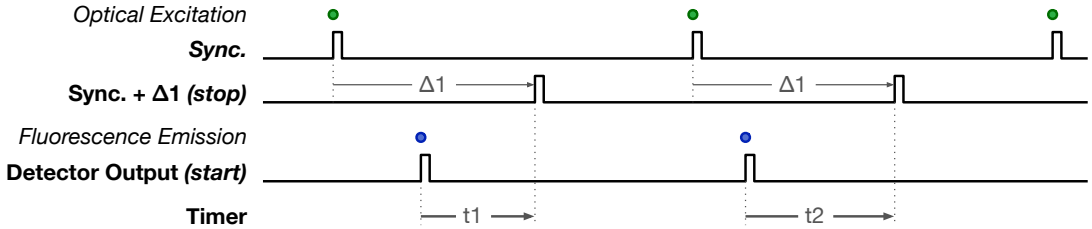


Figure 2.2: *Timing diagram showing reverse start-stop, where the excitation synchronisation is delayed until after fluorescence emission.*

2.2.3 Detection

As the name suggests, TCSPC requires a detector that is sensitive to single incoming photons. Single photon sensitivity requires a very high electrical gain to convert the relatively low energy incident photon flux into a useful electrical output voltage that is observable above any noise floor. TCSPC only requires knowledge that one or more photons have arrived, so a saturating gain that can be approximated as infinity is suitable [62]. The performance of these detectors can be quantified using a number of parameters [18], a subset of which are detailed below. Each of these parameters depend on operating conditions, such as bias voltage and temperature.

- **Sensitivity** is the Volts produced per photon and is given as the photon detection efficiency (PDE) or probability (PDP) and would ideally be 100 % at the emission wavelength.
- **Noise** is given as the dark count rate (DCR) in Hertz at given operating conditions.
- **Transit Time Spread** or **Jitter** is the variation from event to event of the delay between the optical input and the output of the detector's electrical signal. This is typically quoted as the full width at half maximum (FWHM) of the temporal distribution.
- **Dead-time** is the time required to reset (or quench) the detector after detecting a photon. Subsequent photons arriving within this dead-time cannot be distinguished at the output.
- **Afterpulsing** is the likelihood of the detector producing a non-photon induced electrical output pulse directly following and caused by a previous real photon induced event.

By far the most common detector for use with TCSPC is the conventional photomultiplier tube (PMT) and its channel and micro-channel plate (MCP) variants [63, 64]. PMTs operate using a photocathode that creates a small number of electrons when struck by incident photons, caused by the photoelectric effect. These electrons are then directed to a sequence of electrodes, called dynodes, that are held at increasingly higher voltage potentials. The dynodes accelerate the electrons due to the electric field, and hence multiply their number at each stage to produce the required gain. After several stages of multiplication, the electrons strike an anode which produces a fast high current pulse to indicate the arrival of a photon. PMTs have a reasonable PDE of 10 – 40 %; a low DCR, typically below 100 Hz but can be improved by cooling; transit time spreads up to hundreds of picoseconds; dead-times in the nanosecond and sub-nanosecond region; and afterpulsing probability is typically below a few percent, though is very dependent on each individual PMT and its setup and can be as high as tens of percent [18, 65]. Furthermore, PMTs have relatively large active areas, typically mm^2.

However, the electrical output pulses created by a PMT do not have a stable amplitude. A broad distribution of pulse heights (amplitude jitter) is caused by the random amplification mechanism of the detectors, varying light levels, and power supply and temperature variation. This adds the requirement of a discriminator at the output of each PMT in the TCSPC system, as shown in Figure 2.1 [18, 19]. However, a simple leading edge discriminator is not sufficient as the amplitude jitter translates into timing jitter proportional to the pulse rise time. Therefore, constant fraction discriminators (CFDs) are required, which operate on the principle of calculating the difference between the input pulse and a delayed version of itself, then using the zero-cross point which is independent of amplitude to trigger an output pulse. Furthermore, PMTs require a vacuum to operate, which significantly impacts on their reliability, scalability and practical usage lifetime as well as making them unsuitable for operation in high magnetic field environments. Additionally, they require high operating voltages in the kV region and can be permanently damaged if exposed to high light levels, such as direct sunlight.

Avalanche photodiodes (APDs) operating in Geiger mode, or single photon avalanche diodes (SPADs) are becoming more popular due to advances in increased sensitivity, decreased DCR and improved timing jitter [63]. However, they suffer from long dead-times in the tens of nanosecond to microsecond region to keep afterpulsing to a minimum and have very small active areas on the scale of hundreds of square micrometers to achieve optimal jitter and DCR performance. Despite these problems, high-performance SPADs and SPAD architectures have recently been developed in standard CMOS processes, which enables their integration with complex electronic processing. These devices and architectures will be discussed in detail in Sections 2.5.2 and 2.5.3, respectively.

2.2.4 Timing

The main aim of the timer in a TCSPC system is to digitise the time between the excitation synchronisation pulse and the pulse created by the detected photon emission with sufficient resolution so that fluorescence lifetimes down to sub-nanoseconds can be resolved accurately. The time-to-analogue converter (TAC) plus analogue-to-digital converter (ADC), or TAC-ADC method is the most common timing technique for TCSPC [18]. As shown conceptually in Figure 2.3, the TAC operates on the principle that a capacitor (C) is linearly charged, using a constant current source (I), between successive *start* and *stop* pulses. The ADC then converts the voltage generated across the capacitor, which represents time, into a digital signal to be

further processed. Once the ADC has completed its conversion, the charge on the capacitor is grounded, preparing it for the next event. The TAC-ADC method is used for TCSPC due to its time resolution of down to sub-picoseconds, made possible by the use of a 12-bit ADC. However, it does have a number of drawbacks: amplification stages are required between the TAC and ADC, as shown in Figure 2.1, complicating the set up and introducing additional circuit non-linearities; and ADC conversion times are in excess of hundreds of nanoseconds, which significantly affects the available photon throughput, compounding the pile-up problem, as will be introduced in more detail in Section 2.3.

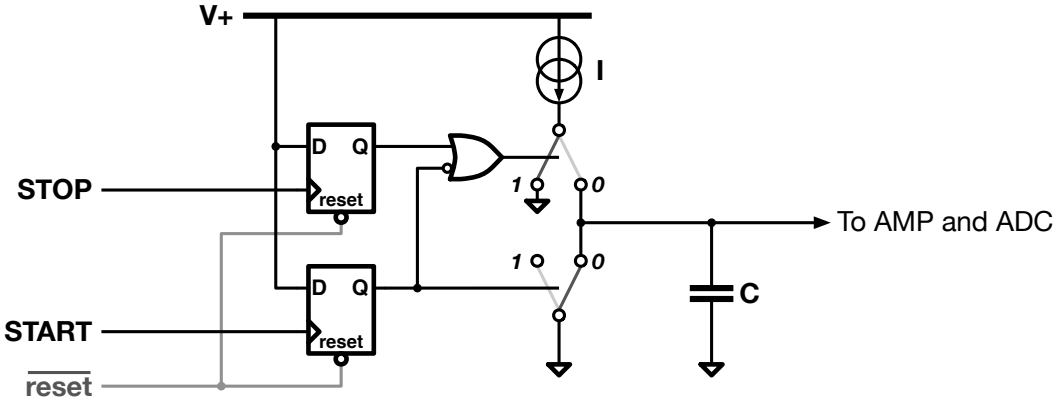


Figure 2.3: *Simplified Time-to-Amplitude Converter (TAC) approach to TCSPC timing. [18]*

The complexity and conversion time limitations of TAC-ADC approaches can be overcome by using direct time-to-digital converter (TDC) techniques – which use the delay through a chain of logic gates for time measurement – at the expense of timing resolution [66, 67]. The most basic TDC principle will input the *start* pulse into an inverter chain and the *stop* pulse is then used to sample the state of the chain into registers, as shown in Figure 2.4. The time resolution of this approach is determined by the minimum gate delay of the inverter in the process technology being used and is typically in the tens to hundreds of picoseconds range. However, this simplistic approach suffers from a number of drawbacks: firstly due to the unary encoding scheme of a linear chain, 4096 inverters plus latch or flip-flop elements are required to reach the same number of time bins as a 12-bit ADC; and secondly the technique is very sensitive to process, voltage and temperature (PVT) variations, which causes instabilities over time and non-linearities due to transistor mismatch. A number of techniques exist to overcome the limitations of the naive delay line approach described above. These will be introduced in Section 2.5.4, where the description of a suitable architecture for embedding on a miniaturised multiple timing channel sensor is presented.

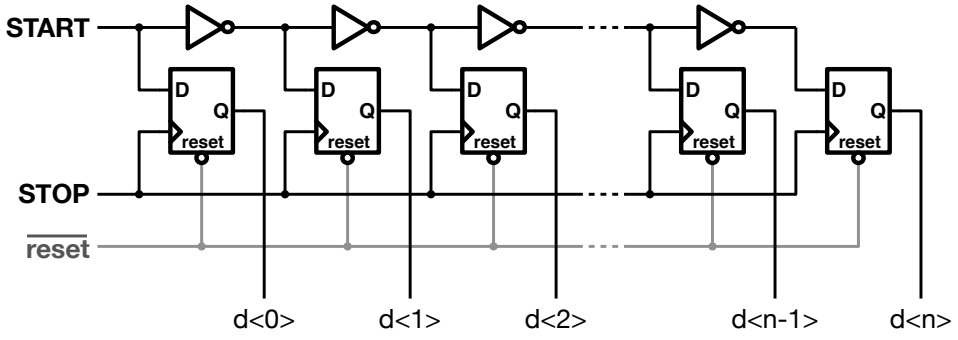


Figure 2.4: *Simplified Time-to-Digital Converter (TDC) approach to TCSPC timing. [18]*

2.2.5 Data Handling

The simplest procedure to deal with digitised TCSPC time-stamp data is to increment a counter held in memory that is addressed by the time-stamp itself [18]. Creating a histogram by binning in this way is an efficient compression technique, requiring only enough memory to hold sufficient counts in each TCSPC bin (e.g. a 12-bit counter for each of 1024 time bins requires only 1.5 kB of memory). This is a popular approach for laser scanning and/or multiple detector TCSPC set ups, however to store a full histogram for each X-Y position (pixel) or detector, memory requirements quickly grow (e.g. 32 channels recording 512×512 images requires 12 GB of memory). Furthermore, it does not provide macro-scale temporal information about when in an experiment each TCSPC event occurred.

The time-tagged time-resolved (TTTR) approach to data handling is a more flexible technique [54, 68]. TTTR works on the principle that each individual photon event is unique and independent, so should be stored as such. The advantage of this is that each event can be saved with *meta* data – such as macro-time (time since the start of the experiment), as well as X-Y position (pixel) and/or detector channel information. However, due to the uncompressed nature of the data, this technique requires a large buffer, operating in a first-in first-out (FIFO) configuration and a high speed data-link to a PC to save the data to memory, process it and/or write it to hard disk. Furthermore, the demands on this approach will be increased if throughput of individual channels is increased by overcoming the pile-up limit.

Neither of these approaches is suitable for low-latency, *real-time* applications where even higher compression using embedded processing is required to reduce the bandwidth requirements [9]. Techniques to achieve this *real-time* calculation will be introduced in Section 2.6.

2.3 TCSPC Pile-Up

2.3.1 Overview

TCSPC has one major drawback – the limited photon throughput available as a consequence of the inability of the hardware to process every detected[1] photon event. This problem is commonly referred to as the TCSPC *pile-up* limit and is caused by a number of hardware constraints. The result of pile-up is that experiments are typically only performed with photon count rates of up to a predetermined limit, the value of which depends on a range of specific experimental requirements such as the acceptable error in lifetime calculation, the speed of the TCSPC apparatus and the values of the lifetimes being measured. Typical limits on the photon count rate range from up to 1 %, 5 % and 10 % [18, 19] of the excitation repetition rate. The reason these numbers are quoted as a fraction of the excitation rate will become clear in the following sub-section, where the causes of pile-up are explored together with the effect each has on TCSPC captured lifetime decay curves. The section finishes with a discussion of techniques currently used to overcome or reduce the effect of TCSPC pile-up in single-channel systems.

2.3.2 Causes of Pile-Up

Classical (Timer) Pile-up

The first form of TCSPC pile-up is the inability to process more than one photon per excitation period and is caused by TCSPC systems relying on a single TAC-ADC to perform the time digitisation. This is sometimes referred to as *classical* [69] or statistical (S-type) [20] TCSPC pile-up, but will also be referred to in this work as timer or timing pile-up. As shown by the conceptual timing diagram in Figure 2.5, the effect of only having a single timer to measure events means that only photon arrivals 1, 4 and 6 can be timed, represented by $t1$, $t2$ and $t3$, whilst photon arrivals 2, 3, 5, 7 and 8 are missed. In this example, the dead-times of both the time conversion and the detector pulses are assumed to be ideal (i.e. zero).

[1] The photon rates discussed from this point on refer to the number of photons that cause the active area of the detector to register an event and is assumed to be directly proportional to the emission intensity.

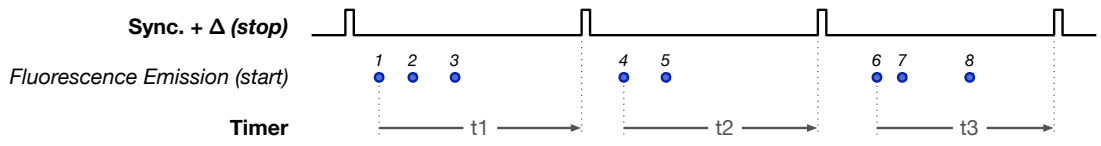


Figure 2.5: *Classical (Timer) TCSPC pile-up.*

Due to the *first* event in any given excitation period being timed without any issues, this form of pile-up causes photon events that arrive *late* in the decay to have a higher probability of being lost or missed, which in turn causes the TCSPC captured decay histogram to appear to distort towards a *shorter* lifetime. This effect can be described analytically by Equation 2.1, which has been adapted from [18], where a proof is provided. In the equation the standard lifetime decay, as given by Equation 1.1 ($I_0 \cdot e^{-t/\tau}$), is multiplied by the probability that the event at time t was the first photon event within the given excitation period. The I_0 term from Equation 1.1 has been dropped for simplicity, so Equation 2.1 describes a decay that is normalised to $I_0 = 1.0$. The result of this equation is shown in the graph in Figure 2.6 for a range of average photon-rates, μ, from 0.01 to 10.0. As expected, the graph shows the decay distorting towards a shorter (normalised) lifetime for increasing μ.

$$I'(t;\mu) = e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)} \tag{2.1}$$

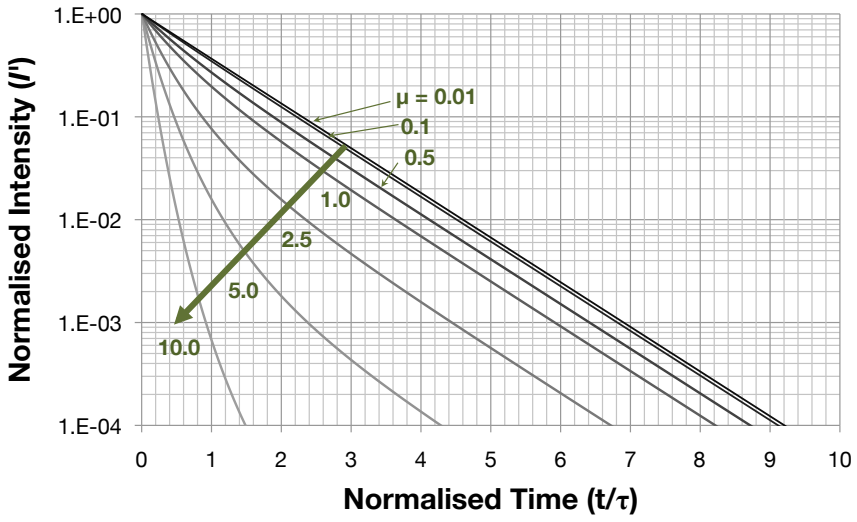


Figure 2.6: *Classical (timer) pile-up – effect of increasing the detected photon rate (μ) on the TCSPC captured decay histogram.*

The average intensity of the normalised fluorescence decay as a function of the excitation period, T, is given by Equation 2.2[2]. This representation of the average intensity ($\bar{I}$) is directly proportional to the number of discrete photons.

$$\bar{I} \quad = \quad \frac{1}{T} \cdot \int_0^T e^{-t/\tau} \, \mathrm{d}t \quad = \quad \frac{1}{T} \cdot \left[-\tau \cdot e^{-t/\tau} \right]_0^T \quad = \quad \frac{\tau \left(1 - e^{-T/\tau} \right)}{T} \tag{2.2}$$

In fact in a typical experiment, the excitation period is likely to be at least 5-10 times longer than the lifetime being measured, so the exponent term ($e^{-T/\tau}$) will approach zero. Equation 2.2 can then be simplified and the average intensity can be approximated by Equation 2.3.

$$\bar{I} \approx \frac{\tau}{T} \qquad \text{for } T > 5\tau \tag{2.3}$$

The same steps can then be taken to calculate the average intensity of the normalised fluorescence decay suffering from classical timer pile-up by using Equation 2.1 in place of the ideal decay, as shown by Equation 2.4. Again assuming $T > 5\tau$, the exponent term ($e^{-T/\tau}$) will approach zero, so the average intensity can be approximated to Equation 2.5.

$$\bar{I}'(\mu) \quad = \quad \frac{1}{T} \cdot \int_0^T e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)} \, \mathrm{d}t \quad = \quad \frac{\tau \left(1 - e^{-\mu\left(1-e^{-T/\tau}\right)} \right)}{T \cdot \mu} \tag{2.4}$$

$$\bar{I}'(\mu) \approx \frac{\tau \left(1 - e^{-\mu} \right)}{T \cdot \mu} \qquad \text{for } T > 5\tau \tag{2.5}$$

The portion of intensity lost – or the probability of missing photons – due to classical TCSPC pile-up can now be defined by subtracting the ratio of Equation 2.5 (the average intensity with pile-up) and Equation 2.3 (the average intensity of the ideal decay) from one, as shown by Equation 2.6.

$$Pr_{missed}(\mu) \quad = \quad 1 - \frac{\bar{I}'}{\bar{I}} \quad = \quad 1 - \left(\frac{1 - e^{-\mu}}{\mu} \right) \tag{2.6}$$

[2] T is assumed to be long enough so that the chance of fluorescence occurring at $t > T$ is negligible.

This equation can also be derived assuming photon events arrive according to a Poisson distribution, with mean μ. The probability of a single photon being processed is equal to the probability of one *or more* photons occurring within an excitation period: $P(X \geq 1)/\mu$. Therefore the probability of *missing* photons due to classical TCSPC pile-up can also be defined by subtracting this from one, as shown in Equation 2.7.

$$Pr_{missed}(\mu) \; = \; 1 - \frac{P(X \geq 1)}{\mu} \; = \; 1 - \left(\frac{1 - P(X = 0)}{\mu} \right) \; = \; 1 - \left(\frac{1 - e^{-\mu}}{\mu} \right) \qquad (2.7)$$

Plotting this result against μ yields the graph shown in Figure 2.7, which highlights how much of an impact this form of pile-up has on the number of missed photons: increasing μ from 0.1 to 0.5 and 1.0 causes the number of missed photons to increase from 5 % to 21 % and 37 %, respectively, and 90 % of photons are missed for $\mu \; = \; 10.0$.

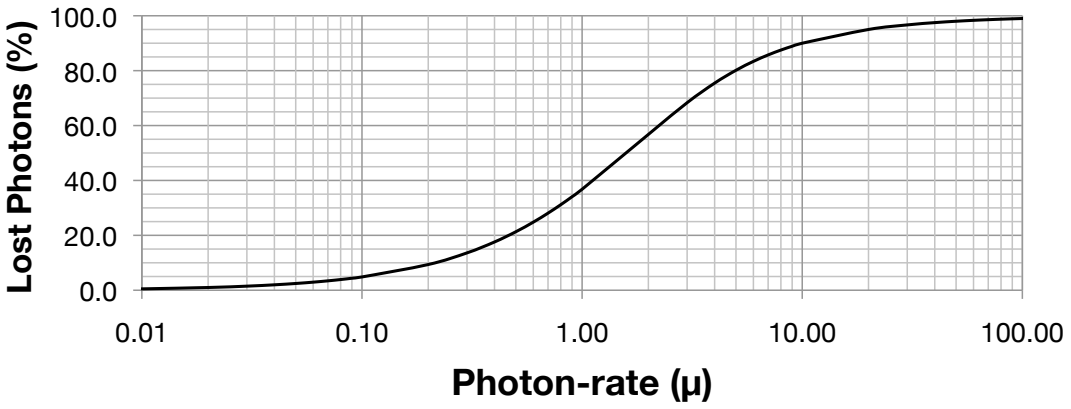


Figure 2.7: *Relationship between the photon-rate (μ) and the percentage of photons lost to classical TCSPC timer pile-up.*

Before the introduction of high-repetition rate pulsed excitation sources, this form of pile-up was the most limiting from a practical experimentation perspective, as system dead-times (which will be introduced in the following two sub-sections) were orders of magnitude shorter than the excitation periods. As the repetition rates of excitation sources have become faster, these system dead-times become more and more of an issue. The graphs in Figure 2.6 and 2.7 therefore represent the absolute *best* case achievable by a conventional single-channel TCSPC system and as the following sub-sections will describe, in practice TCSPC pile-up has a significantly worse effect on the captured decay histograms.

Non-Extending Dead-Time (Conversion) Pile-up

In addition to a TCSPC system only being able to time at most one photon event in every excitation cycle, the electronic components used to perform this measurement require some processing (or conversion) time before they are ready for a subsequent photon arrival. During this time, the TCSPC timing hardware is assumed to be dead and cannot process any further events. The dead-time can arise from a number of components in the TCSPC system, such as the discriminator(s), TAC, amplifier, ADC and data processing or memory accesses. This form of pile-up has been referred to as electronic (E-type) pile-up and can be classified as having a *non-extending* dead-time [20]. The duration of a non-extended dead-time is not altered by the arrival of any photons within the dead-time. This form of pile-up will be referred to as conversion pile-up or timer dead-time in the remainder of this thesis.

As can be seen in the conceptual timing diagram in Figure 2.8, the conversion time compounds the problems already apparent with timer pile-up. In this simplified case, the hardware is only capable of timing photon events 1 and 7, represented by $t1$ and $t2$, whilst 2, 3 and 8 are missed due to classical timer pile-up and 4, 5 and 6 are missed due to conversion dead-time pile-up caused by $t1$ ($t1$ *Conversion Time*). As the conversion time is non-extended, photons 4 & 5 do not cause the $t1$ *Conversion Time* to change, and furthermore the $t1$ *Conversion Time* will be approximately equal to those of $t2, t3, ...tn$.

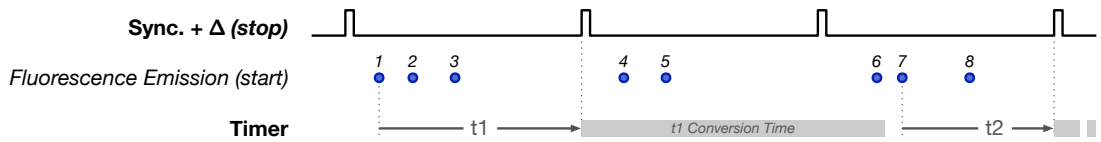


Figure 2.8: *Non-Extending Dead-Time (Conversion) TCSPC pile-up.*

Excitation repetition rates have been increasing at a much higher rate than improvements to TCSPC conversion times. Modern commercially available TAC-ADC TCSPC system architectures have conversion dead-times in excess of 100 ns [18], which in itself is a full period of a 10 MHz laser or ten periods of a fast 100 MHz laser. In fact, when the conversion dead-time is greater than the excitation period, it no longer makes sense to describe the maximum photon-rate as a function of the excitation rate, but rather as a function of the TCSPC system conversion dead-time. However, in the interest of consistency and comparison the discussions and results presented throughout this thesis will still refer to photon-rates as a function of the excitation repetition rate, regardless of conversion time.

Extending Dead-Time (Detector) Pile-up

The final form of TCSPC pile-up is caused by the finite output pulse-width of the single-photon detector. This pulse-width is also referred to as electronic (E-type) dead-time [20] and if subsequent photons arrive within this dead-time then they cannot be distinguished by the timing electronics. Unlike TCSPC conversion dead-time however, detector dead-time is extending, meaning that if a photon event arrives within the dead-time of the previous event, then the detector will remain dead for at least another full dead-time. Because of this, at very high photon rates, it is possible for detector dead-time to be saturating, meaning only the first arrival can be distinguished.

Detector dead-time can be thought of as pulse collision, or pulse overlap and an example of it is shown in Figure 2.9, where the detector output only creates an electrical rising edge for photon events 1, 4, 6 & 8, whilst events 2, 3, 5 & 7 are indistinguishable. The extendability of detector dead-time is also apparent in the figure, where photon event 2 causes the output to remain high long enough to affect photon event 3. If timer and conversion dead-time are also considered, then only photon events 1 and 8 could be timed ($t1$ and $t4$).

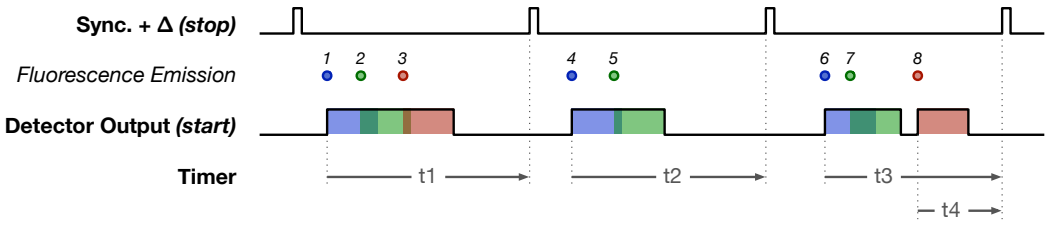


Figure 2.9: *Extending Dead-Time (Detector) TCSPC pile-up.*

In traditional single channel TCSPC systems, detector pile-up is not a major issue as the dead-times of PMTs are in the region of sub-nanoseconds to nanoseconds and so very high photon rates over an order of magnitude greater than the pile-up limit (in excess of the excitation rate) are necessary before its effects are noticeable [69]. However, the use of SPAD detectors for TCSPC, which have much longer dead-times in the region of tens to hundreds of nanoseconds, will have a noticeable effect at lower photon rates and even more of an effect as photon rates are increased by overcoming the classic timer and conversion dead-time pile-up limits. SPAD dead-time considerations will be reviewed in Section 2.5.2, whilst an in depth study of the detector pile-up effect will be given throughout Chapter 3.

2.3.3 Techniques to Overcome Single-Channel TCSPC Pile-Up

A number of techniques have been developed to increase photon throughput in fluorescence lifetime experimentation by reducing the effects of TCSPC pile-up or overcoming it completely. An outline of some of these techniques will be introduced below, with a critical discussion of the limitations and disadvantages of each. It should be noted that although some of the techniques described below are readily achievable, due to their individual complexities and/or inefficiencies, they are not all in widespread use and the primary method of choice for dealing with single channel TCSPC pile-up is to simply operate the experiment at photon count rates below 1 – 10 % of the excitation rate. Multi-module TCSPC approaches to reducing the effect of the pile-up limit, which are more popular, will be introduced in Section 2.4.

Post-Processed Correction

An analytical technique was first proposed by Coates [70] in 1968 to correct the captured TCSPC decay histogram after the completion of the experiment. The technique required only the captured TCSPC data set (histogram) and knowledge of the total number of excitation pulses used in the experiment. The simplicity of this approach is achieved by not requiring any prior knowledge of the photon rate, which is difficult to measure and quantify due to its variability. Improvements to this analytical approach were made by Walker [71] in 2002, who describes an iterative algorithm to allow the correction to operate under variable excitation energy, which was not considered by Coates. The major disadvantage of this technique is the requirement of an additional post-experiment computation step, which can be very time consuming for large data-sets to correct the fluorescence lifetime decay curve(s) before they are analysed further. Furthermore, due to the additional analysis step it is less suitable for applications requiring *real-time* data processing.

Inhibit Circuit

Following on from the work by Coates, a hardware approach to overcoming the TCSPC pile-up limit was first introduced by Davis and King [72] in 1969 and then by Williams and Sandle [73] in 1970. If the TCSPC hardware only saves a time-stamp to memory when the photon event that created it was the *only* event in the given excitation period, then the captured decay histogram will appear completely unaffected by classical timer pile-up. This technique is achieved in practice by adding an inhibit circuit to the standard TCSPC set up that performs standard photon

counting of the output from the fluorescence detector. If the inhibit circuit counts more than one photon in an excitation period, it sends a signal to one of the components in the processing block (TAC, ADC or memory access) to inhibit it from writing the data of the timed photon event to memory. The counter of the inhibit circuit must be reset at the beginning of each excitation period.

The main advantages of this approach are that it completely solves the classical timer pile-up limitation and the technique is user-friendly as the pile-up photons are removed at source in real-time, requiring no further analysis of the captured decay histograms. However, this is achieved at the expense of a reduction in photon collection efficiency as it has the effect of decreasing the number of *processed* photon events as the number of *detected* photons is increased. The probability of *processing* a photon with the inhibit circuit can be described using the Poisson distribution to calculate the probability of *exactly* one photon arriving within an excitation period ($P(X = 1)/\mu$). Subtracting this from one gives the probability of missing photons, as given by Equation 2.8. Plotting this as a function of μ results in the graph in Figure 2.10, which shows that more photons are missed using the inhibit circuit than due to classic TCSPC pile-up, as expected (see Figure 2.7). Increasing μ from 0.1 to 1.0 and 5.0 causes the number of missed photons to increase from 10 % to 63 % and 99 %, respectively. Furthermore, the technique adds an extra hardware component to the TCSPC set up and the inhibit circuit itself is complex, requiring that its own dead-time is negligible.

$$Pr_{inhibited}(\mu) = 1 - \frac{P(X = 1)}{\mu} = 1 - e^{-\mu} \tag{2.8}$$

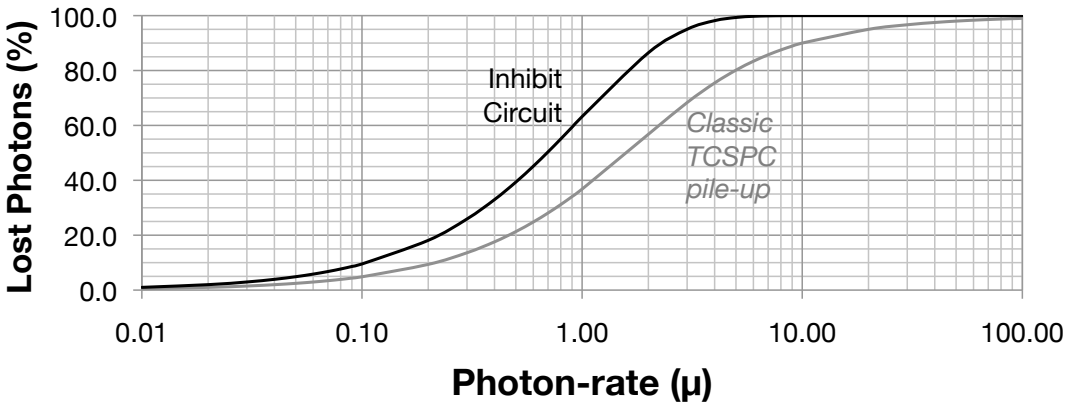


Figure 2.10: *The percentage of photons lost using the inhibit circuit as a function of μ.*

Continuous Time Forward Start-Stop

A completely new approach to the timing architecture for TCSPC was introduced by Wahl *et al.* [74] in 2007, where the concept behind a commercially available instrument was described (PicoHarp 300, PicoQuant). The instrument uses a pair of TDCs that are operated in a continuous free-running mode, where at any given moment they represent a picosecond accurate time from the beginning of the experiment. This is made possible by the combination of long time range TDCs together with the addition of an overflow flag introduced into a TTTR data stream to tell when they have wrapped around. The technique is shown conceptually in Figure 2.11, where the *Sync. Timer* and *Emission Timer* represent the two free-running TDCs. The times from the TDCs are independently sampled by incoming excitation synchronisation pulses (s*) and photon events (e*), respectively. The TCSPC time-stamps are then calculated in real-time, using an FPGA to subtract the most recent time-stamp sampled by the synchronisation TDC from the current time-stamp sampled by the fluorescence detection TDC, as shown at the bottom of the figure. This effectively means the instrument operates in a *forward* start-stop mode, which minimises the effect of any conversion dead-time. As the timing of photon events and synchronisation pulses are independent, there is no limit to how many photon events could be processed during each excitation period. This is shown in the figure, where photon events 3 and 8 – which would normally remain unprocessed due to classical timer pile-up – are processed as the conversion time of the previous events (1 and 6 in this case) has completed.

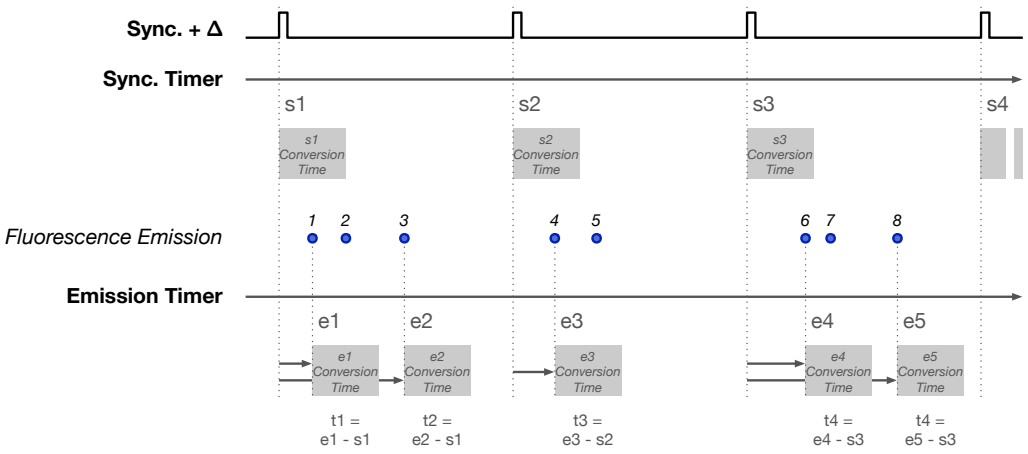


Figure 2.11: *Continuous Time Forward Start-Stop technique to TCSPC.*

One of the difficulties with this approach is that the two TDCs must be kept in synchronisation with each other, which has been achieved in this instance by distributing a common quartz reference oscillator clock to each TDC. However the picosecond time resolution required is faster than can be achieved using only this clock, so time interpolation is necessary. This is the primary source of a conversion dead-time of 90 ns in this implementation. So although the technique is ideally suited for overcoming the classical timer pile-up limitation, conversion dead-time pile-up is still an issue. This is highlighted in Figure 2.11, where photon arrivals 2, 3, 5, 7 and 8 are all missed due to conversion dead-time pile-up of a previous event. Furthermore, the conversion dead-time also affects synchronisation, as shown by *s* Conversion Time*, so if the excitation rate is above 10 MHz, the electrical synchronisation pulses must be divided down below 10 MHz, resulting in a captured histogram containing multiple decays that must be merged post-experiment. The approach however is very well suited to overcoming the classical timer pile-up limit for experiments requiring the measurement of long lifetime fluorophores, where excitation repetition rates many times slower than the conversion dead-time are required.

Faster Hardware

An approach to minimise the effects of TCSPC pile-up (rather than attempt to overcome it) has been presented by McLoskey *et al.* [75] (of Horiba Jobin Yvon) as recently as 2011. This work does not introduce any *additional* hardware or processing requirements, but to the best of the author's knowledge represents the state-of-the-art in both high-speed excitation sources and short conversion dead-times for TCSPC. A semiconductor impulse generator is presented that is capable of driving a diode excitation source to create 64 ps optical pulses at a frequency of up to 100 MHz. Furthermore, the timing circuitry used has a low conversion dead-time below 10 ns, which matches the period of excitation and represents an order of magnitude improvement over conventional TCSPC timing electronics [18]. The combination of fast excitation rates and low conversion dead-times allows photon throughput rates, in real terms, of up to 10 Mcps.

However, the approach of simply increasing the excitation repetition rate to minimise classical timer TCSPC pile-up and increase photon throughput in real terms is not suitable for all experiments. The excitation repetition rate of 100 MHz that is presented only allows the resolvability of fluorescent molecules whose lifetimes are significantly below the excitation period, or under $\approx 1 - 2$ ns in this case. To measure longer lifetimes, the excitation repetition rate must be slowed down, which in turn reduces the photon throughput rate.

2.4 Integrated Multi-module TCSPC Architectures

2.4.1 Overview

In his book [18], Becker states that *the only solution to the count rate problem is multi-module operation*, where several detectors are connected to a number of independent TCSPC modules to increase the counting capability. However, this approach further increases the cost, size and complexity of an already expensive, large and complex experimental technique. Nonetheless, the approaches and architectures of these multi-module systems can provide an insight to direct the design of an integrated sensor, where each *module* can be implemented as a block within a miniaturised CMOS architecture. This section will therefore review different multi-module architectures and discuss their suitability for both overcoming the TCSPC pile-up limit and for integration in a miniaturised CMOS sensor. In addition to the number of detectors (N_D) and number of or timers (N_T), the concept of a number of channels (N_C) will be introduced to help describe all possible architectures and sub-architectures.

2.4.2 Multiple Detectors, Single Timer

The most common multi-module architecture is a multiple detector arrangement ($N_D > 1$) with a single timer ($N_T = 1$), as shown by the diagram in Figure 2.12. In this case, as there is only one timer, the number of channels (N_C) is also one. This multi-module approach is typically used for detecting the fluorescence lifetime at a small number of different wavelengths ($N_D \leq 4$) simultaneously by using different optical filters for each detector [76, 77]. The general architecture has also been used for integrated solid-state SPAD image sensors, such as: [78], where an entire 64×64 array of detectors use a single embedded TDC; and [79], where the approach is used as a sub-architecture to partition a 128×128 array of detectors into groups of 4×128 per TDC.

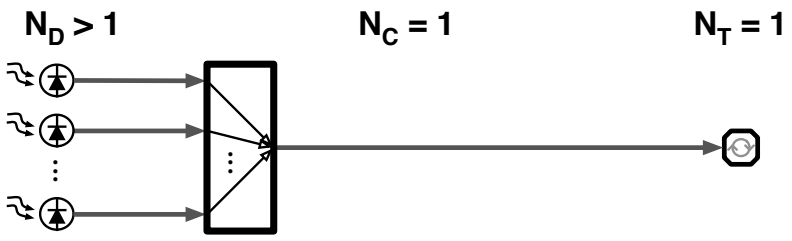


Figure 2.12: *Multiple detector, single timing channel TCSPC architecture.*

The implementations introduced above highlight two different channel recombination techniques: in the first instance, an encoding router is required to inform the TCSPC data processing block which detector a photon event originated from, so the time-stamp can be saved to the correct histogram memory array or labelled appropriately for TTTR acquisition; and in the second instance, the detectors are time-interleaved with the TDC, so only one detector has access to the TDC at any given time. Furthermore, a third recombination possibility exists for multiple-element, single output detectors such as silicon photomultipliers (SiPMs) that will be introduced in Section 2.5.3.

As expected however, this multi-module architecture does not help improve photon throughput. In fact, it makes it worse per detector due to the single timing channel being the source of the most limiting pile-up (timing and conversion). Nevertheless, the structure introduced is commonly used so is important to understand for developing an improved TCSPC architecture.

2.4.3 Multiple Detectors, Multiple Timers

A common technique to overcome the worsened pile-up performance in multiple detector arrangements is to provide a timer for each individual detector, as discussed in [76, 77]. Such architectures are also described in [80–82] and can be used as multiple independent detection channels for multi-wavelength detection. More importantly however, if the detectors can be positioned physically close enough to each other to make them appear as one detector from an optical perspective, then the results from each channel can be combined to overcome the pile-up limit by the same order as the number of detectors and timing channels. The general architecture is shown in Figure 2.13, where in this case there are now an equal number of channels (N_C) as detectors (N_D) and timers (N_T). However, such set ups are very expensive due to the requirements of the multiple individual detectors, timing channels and processing power to keep up with high data rates.

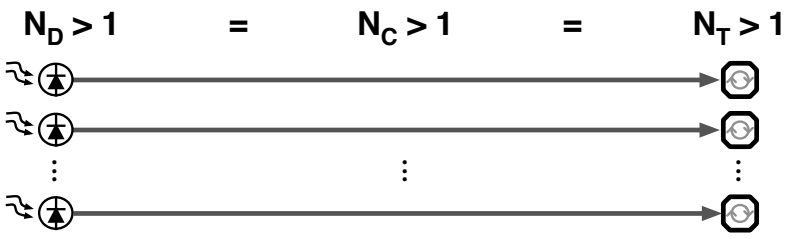


Figure 2.13: *Multiple detector, multiple timing channel TCSPC architecture.*

This architecture is also the basis for recent developments in completely parallel solid-state TCSPC arrays, where a TDC is embedded within each pixel directly beside a SPAD detector. This has been achieved using a standard CMOS imaging process for both 32×32 [12] and 160×128 [22] arrays. In theory, by combining the data from multiple *pixels* together, this multi-module architecture provides the ideal solution to overcoming the TCSPC pile-up limit. However it relies on the fluorescence emission being efficiently distributed onto the active areas of all detectors, which is not a trivial task. In the implementations referenced, each *pixel* is 50×50 µm but is largely occupied by the TDC, whilst the SPAD has a diameter of only 6.7 µm. This creates an optical fill-factor of < 2 %, as highlighted in the photo-micrograph in Figure 2.14. The addition of optical concentrators on the surface of the chip to recover the fill-factor losses has proved unsuccessful thus far [83].

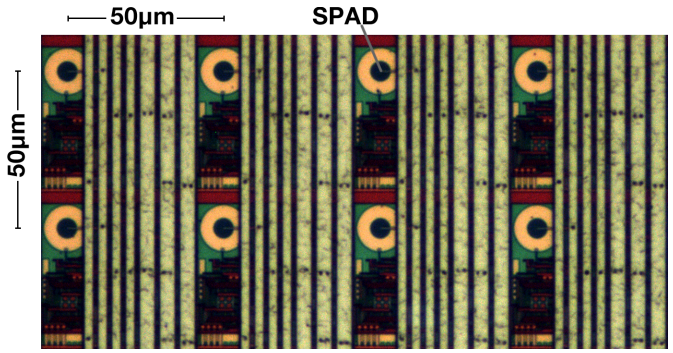


Figure 2.14: *A 2×4 sub-array of pixels from a 32×32 TCSPC imaging array. [12]*

The use of multiple beam-let arrays to match the detector dimensions, as implemented using diffractive optical elements (DOEs) [84] or spatial light modulators (SLMs) [4, 5, 7, 8] provides an option to overcome the fill-factor limitation of these devices. However, neither of these techniques scales well beyond a few tens of individual excitation spots due to excitation power limitations. Furthermore, they are notoriously time consuming, complex and sensitive to set up and operate effectively, even with automated procedures. In addition to optical limitations, these devices are severely limited by data transmission bottlenecks due to the sheer number of parallel TCSPC channels and the limited number of data I/O pads available with the CMOS implementation. As an example, each pixel in the 32×32 array can produce a TCSPC timestamp at up to 1 MHz, so the entire array can produce time stamps at over 1 GHz [12], which at 10-bits per time stamp corresponds to a data throughput of over 10 Gbps. The larger array produces as much as five times this data [22], so clearly some form of compression or signal processing is required to reduce this significantly.

2.4.4 Single Detector, Multiple Timers

A much less common multi-module TCSPC architecture is that of a single detector ($N_D = 1$) with multiple timers ($N_T > 1$), as shown in the diagram in Figure 2.15. The continuous time TDC approach to overcoming TCSPC pile-up as introduced in Section 2.3.3 [74], can be classed as an example of this form of architecture, though it is a very specific case where the second timer is only used to time the excitation synchronisation pulse. To the best of the authors knowledge, the more generalised architecture has only been introduced by O'Connor and Phillips [19], where they describe a multiple channel TAC to allow several photons to be timed during each excitation period. Details of the implementation and a critical discussion of the performance characteristics are limited, however it is claimed to allow a count rate of 50 % of the excitation rate for a 100 ns lifetime fluorophore.

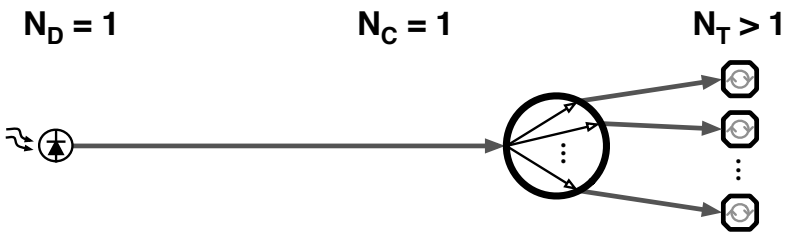


Figure 2.15: *Single detector, multiple timing channel TCSPC architecture.*

Providing there are a sufficient number of timers, this architecture is capable of completely removing classical timer and conversion pile-up. However despite leaving only detector dead-time pile-up, the approach is not commonly used. This is partly due to the dead-times of detectors being in the nanosecond (PMTs) to tens of nanosecond region (SPADs), with only fast PMTs and MCP-PMTs being capable of sub-nanosecond pulses. These values are on a similar scale to those of the lifetimes typically being measured[3], so in most cases only one timer is utilised per excitation period, rendering the additional timers in the architecture redundant. Furthermore, the individual timers must be synchronised with each other and have minimal mismatch which is not a trivial task and will be discussed in detail in Section 2.5.5. Additionally, the high speed asynchronous nature of photon event arrivals creates a complexity in the design of the routing/distribution element within the architecture, which is represented by the large circle in Figure 2.15.

[3] See Tables 1.1 and 1.2 in Section 1.2.4

As discussed, the detector dead-time pile-up problem is worse for CMOS SPAD detectors, which have typical dead-times of 10 – 100 ns. Section 2.5.3 will describe a technique that can be used to minimise this problem by using a single output, multiple-element detector, more commonly referred to as a silicon photomultiplier (SiPM). From a system point of view, this architecture is best described as a single detector as the pulses from the individual detection elements are combined without any positional information. Such an architecture is shown in Figure 2.16, where although there are multiple individual detectors ($N_D > 1$), they are combined into a single output ($N_C = 1$) before being distributed to the timers. This approach overcomes the area and fill-factor constraints of the multiple detector, multiple timer architecture in Section 2.4.3, as the detectors and timers are independent and can be physically separate from each other. This allows the detection elements to be positioned closer together to improve fill-factor. If the detector dead-time issue can be resolved, this architecture represents the most promising solution to both miniaturisation and overcoming the TCSPC pile-up limit.

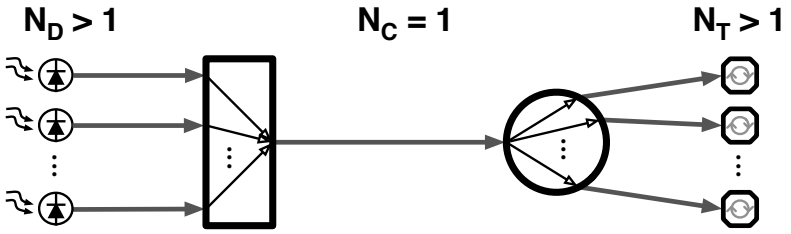


Figure 2.16: *Single channel, multiple-element detector, multiple timing channel architecture.*

2.5 CMOS Technologies

2.5.1 Overview

The primary goal of this thesis is to design, manufacture and test a miniaturised CMOS time domain fluorescence lifetime sensor capable of overcoming the TCSPC pile-up limit. Implementation in a standard CMOS process offers low manufacturing costs for high volume production. For this research, it is necessary to understand the current state of the art technology available to perform both high photon throughput single photon sensitive detection and low dead-time, picosecond accurate time conversion. This section will therefore introduce background, theory and standard CMOS compatible implementations of: single photon avalanche diodes (SPADs), silicon photomultipliers (SiPMs), time to digital converters (TDCs), and finally time-interleaved (TI) converter architectures.

2.5.2 Single Photon Avalanche Diodes (SPADs)

A single photon avalanche diode (SPAD) is a specific form of photodiode structure that is biased beyond its breakdown voltage to provide an apparent infinite optical gain. One would expect that biasing a diode beyond its breakdown voltage would cause an immediate avalanche process to occur, rendering such a device useless. However, the avalanche process can only occur when there are free charge carriers in the depletion region [85]. Therefore, assuming complete darkness and neglecting thermal, tunnelling or defect induced carriers, a diode can be held beyond its breakdown voltage with no current flow. Only when a photon strikes the device is an electron-hole pair created to initiate a self-sustaining avalanche process. This operation is highlighted in the exaggerated IV-curve shown in Figure 2.17, where V_{BD} and V_{OP} represent the SPAD breakdown voltage and the total bias across the SPAD, respectively. As shown in the figure, a quenching process is necessary in order to reset the device back to its steady state [86].

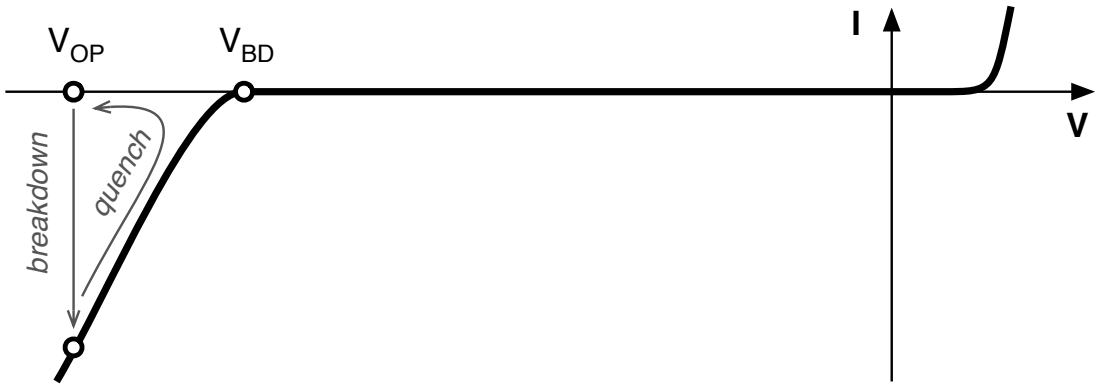


Figure 2.17: *Conceptual diagram of a SPAD IV-curve, showing breakdown and quenching.*

Although quenching can be performed in a number of ways, it is typically performed passively using the circuit shown in Figure 2.18, where the SPAD anode is connected to a negative voltage (V_{BD} – approximately equal to the breakdown voltage of the diode) and the cathode is biased to a positive voltage (V_{EB}) via a quench resistor (R_{quench}). In its steady state, the SPAD has a total reverse bias voltage (V_{OP}) equal to the difference between V_{EB} and V_{BD}. When an avalanche process occurs, the current that flows through the SPAD causes a voltage drop across R_{quench} and the voltage across the SPAD is subsequently reduced. Once the voltage across the SPAD has dropped below its reverse breakdown voltage, the avalanche process is stopped and the device is returned to its steady state. An inverter detects the voltage drop across R_{quench} and provides a digital output pulse to signify the arrival of a photon. The values of V_{BD}, V_{EB} and

R_{quench} determine the rate of the voltage change and hence control the dead-time of the SPAD, which is highlighted in the figure. The output pulse contains no amplitude information, so it is necessary to count the number of pulses within a given time period to determine intensity. This gives rise to SPADs commonly being referred to as Geiger-mode devices.

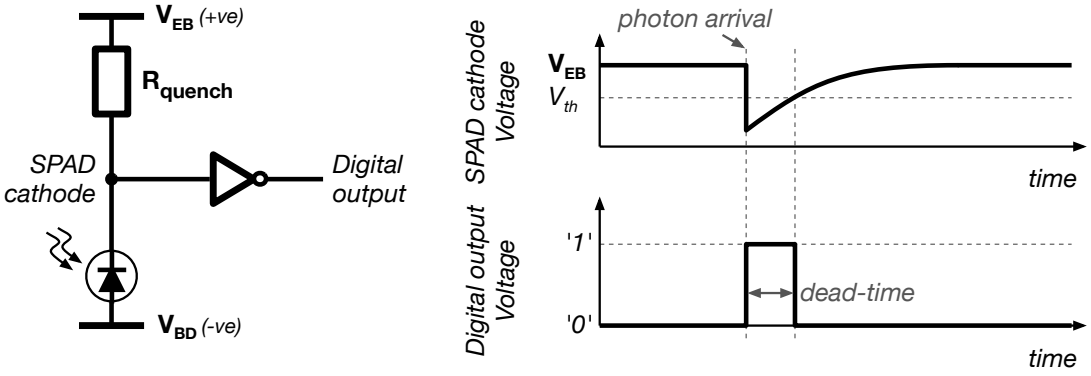


Figure 2.18: *Operation of SPAD and passive quench.*

High performance SPADs have existed for a number of years in non- [87, 88] and custom- [89] CMOS processes. However, due to their non-compatibility with standard CMOS, such devices incur high manufacturing costs and prove difficult to integrate with additional circuitry without complex, bump-bonded, two-chip solutions [90]. Early attempts to integrate SPAD devices in standard CMOS processes produced devices with high dark count rates (DCR) of hundreds of kilohertz [91, 92] and even megahertz [93] as well as relatively low peak photon detection efficiency (PDE) of 5 % at 450 nm [94] and 14 % at 670 nm [95]. However a structure has been developed recently [59] that has significantly improved performance characteristics in these areas. The device has been manufactured and proven in array format [12], which has provided characterisation results across a large sample distribution[4]. The results from these arrays show that 80 % of the population have a sub-100 Hz DCR, caused by thermal generation, whilst the remaining 20 % have an increasing DCR of up to 100 kHz caused primarily by defects in the silicon lattice. Furthermore, the device has a 28 % peak PDP at 500 nm (blue) and sub-200 ps timing jitter, all of which make it suitable for TCSPC experimentation. As explained in Section 1.4, this SPAD technology as well as the 130 nm imaging process it was developed in were made available for use within this research.

[4]See Appendix A.1 for a graphical representation of these results.

The cross section of this SPAD structure is shown in Figure 2.19. Three specific implementation details in particular provide improved performance characteristics over the previously introduced work. Firstly, the diode junction is created between *P-well* and *deep N-well*, rather than the more commonly adopted *P+ / N-well* structures [91–95], as highlighted by the dotted line in the figure. This has the effect of creating a lower electric field multiplication region, widening the wavelength response and reducing tunnelling, which in turn improves DCR yield. It is achieved at the cost of increasing both the breakdown voltage (V_{BD}) and the timing jitter. However, both of these trade-offs are acceptable as voltage can be supplied externally and the timing jitter of 200ps has been proven to be suitable for fluorescence lifetime experimentation [10, 11, 96]. Secondly, a *virtual* guard-ring is introduced, consisting of an area of *deep N-well* surrounding the multiplication region with no additional upper implants. In this area, the *deep N-well* is termed *retrograde* as it has a diminishing doping concentration towards the silicon surface. This ensures a planar breakdown region, as the sides of the *P-well / deep N-well* interface experience a lower electric field – and hence larger breakdown voltage – than that of the main *P-well / deep N-well* interface. Shallow trench isolation (STI) – which is an isolation structure introduced for deep sub-micron processes, where the effects of high doping concentrations requires a physical barrier between diodes and transistors – is used to further isolate the device. Finally, the optimised optical stack (not shown) of the imaging process used for manufacture, consisting of only 4 metal layers in a more advanced (90 nm) process and an improved passivation layer [97] enables the increase in PDP over previous implementations where 6 or more metal layers are used in less advanced processes with standard passivation.

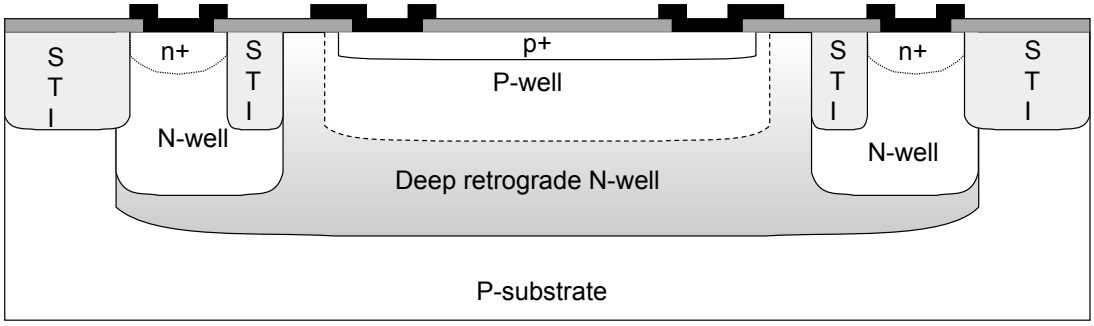


Figure 2.19: *Cross section of 130 nm CMOS SPAD device structure. [59]*

However, these improvements in CMOS SPAD performance come at a cost, in addition to dead-times being orders of magnitude longer than typical discrete PMTs, active areas are also orders of magnitude smaller. The dead-times of SPADs are proportional to the quenching time and are typically in the tens to hundreds of nanosecond region. The long quenching time increases the probability that trapped charge carriers remaining in the multiplication region are completely removed. If any of these *traps* remain filled after quenching is complete, they can cause a secondary (non-photon induced) avalanche, referred to as afterpulsing (see Section 2.2.3). The main reason for long dead-times is therefore to reduce the afterpulsing probability. The dead-time can be reduced by using an active quenching circuit in place of the passive one described. Such a circuit operates by sensing the onset of avalanche, then simultaneously producing a CMOS compatible output whilst quickly reducing the SPAD bias voltage below breakdown before returning it to normal ready for the next photon [98]. The rapid reduction of the bias voltage triggered by the onset of avalanche means the device never fully breaks down, which reduces the probability of traps being filled and therefore reduces the afterpulsing probability. The effect of decreasing SPAD dead-time is highlighted in [99], where passive 30 ns, active 30 ns and active 5.4 ns dead-time SPAD circuits are shown to be capable of 12.3 MHz, 32.9 MHz and 185 MHz maximum photon throughput, for afterpulsing probabilities of 0.57 %, 0.43 % and 1.3 %, respectively. However, active quenching circuits require more transistors – and hence more silicon area – to implement and must be placed in close proximity to the SPAD device for optimal operation.

The active areas of CMOS SPAD devices are typically tens to hundreds of square micrometers, which is necessary for the improved jitter and DCR performance [100]. The improved jitter performance is achieved in smaller devices due to a faster breakdown believed to be caused by a reduced lateral avalanche build-up time. The DCR yield improvement in smaller devices is attributed to a lower probability of the device containing a defect in the silicon lattice. Furthermore, the shape of the SPAD device also affects performance, with studies of square, octagonal, circular, rounded corner square and *Fermat* devices showing that circular devices are optimal. This is due to their lack of corner regions, where defects caused by lattice stresses are more likely to result during the manufacturing process. Both dead-time and active area considerations can be significantly improved upon using multiple element, combined output architectures, typically referred to as silicon photomultipliers (SiPMs). Such architectures will be described in the following section together with details of circuit techniques capable of overcoming these drawbacks.

2.5.3 Silicon Photomultipliers (SiPMs)

Research has been undertaken in recent years in implementing compact SPAD arrays as image sensors – where each individual SPAD can be independently accessed and its output read [78, 101] – and also as silicon photomultipliers (SiPMs) – where the outputs from the multiple SPADs in the array are combined to produce a single detector output [102, 103]. The former have typically been used in ranging and fluorescence lifetime applications whilst the latter have typically been used in high-energy medical imaging, such as positron emission tomography (PET), where the devices are integrated with a scintillator to convert the gamma radiation into visible light detectable by the sensor [104, 105]. There are however no reasons against using a SiPM to perform standard *single detector* TCSPC for fluorescence lifetime [106] and in fact is an ideal solution to solving the issue of small SPAD active areas as the total active area of a SiPM is the sum of active areas of all elements within it.

Conventional SiPMs, which can also be classified as *analogue* SiPMs, do not include the inverter shown in Figure 2.18. Instead they connect all of the SPADs in parallel, as shown in Figure 2.20a, causing a stepped *current* output whose amplitude is dependent on the number of simultaneous avalanche processes occurring at any given time. This approach is similar to the expected operation of a standard PMT, however the device requires an analogue to digital converter (ADC) at the output and the architecture suffers from increased noise levels and comparably poor timing and jitter performance. Digital SiPMs have come into prominence in recent years [107, 108], where the ADC is local to each SPAD in the form of a standard inverter. The output of the multiple inverters are then passed through a logical OR tree to produce a single digital output, as shown in Figure 2.20b. This approach isolates noise and the logic cells re-time the pulse at each level to ensure improved jitter performance. For TCSPC applications, where a stepped output is unnecessary and time resolution is critical, clearly the digital SiPM architecture is more suitable.

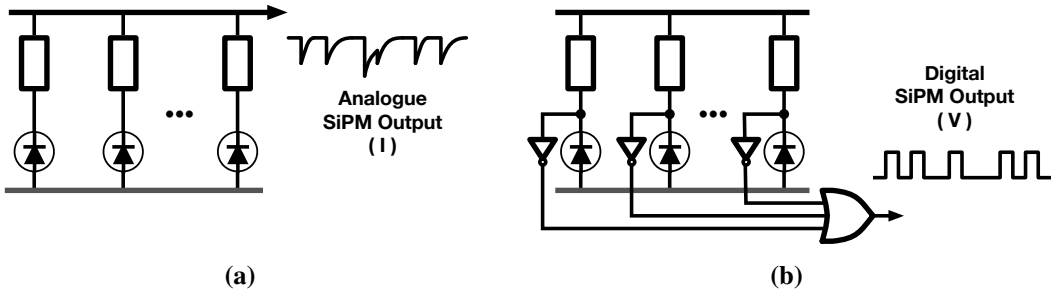


Figure 2.20: (a) *Conventional analogue and* **(b)** *digital SiPM architectures. [108]*

However in their most basic form, digital SiPMs suffer from three primary drawbacks. Firstly, the DCR measured on the output of the logic tree is the sum of the DCR of all elements within it. Therefore, even if only a few percent of devices have tens or hundreds of kilohertz DCR, this can render the device unusable. Secondly, although the SiPM architecture is capable of overcoming the small active area issue of SPADs, the result is to introduce a new problem of reduced fill-factor caused by: the area required by the guard ring structure; the area lost due to the circular shape of the device; the spacing rules of the process; and the requirement of a large-area, high-value poly resistor for R_{quench} located in close proximity to the SPAD. Finally, if one SPAD fires, it will lock-up the output of the logic tree for the duration of its dead-time. As SPAD dead-times are comparatively long when passively quenched, this can be a major issue leading to poor performance with regards to detector pile-up. Fortunately techniques exist to minimise the effect of each of these issues and will be discussed individually below.

The DCR problem of SiPMs can be addressed by producing an enable signal for each element. Performing a *self-test* by enabling SPADs one by one, it is possible to build a DCR map of the device that can be used to disable individual elements that are deemed to be above an acceptable limit or threshold [107, 109]. This technique for SiPMs is almost identical to approaches for correcting defective pixels in CMOS image sensors [110, 111]. The major difference is that the defective (high-DCR) elements in a SiPM are simply ignored rather than being corrected as is necessary for pixels in an image sensor. Furthermore, the technique allows the active area, and hence sensitivity, of the SiPM to be adjusted by enabling a specified number of adjacent elements.

Fill-factor limitations can be improved for a reduction in performance of other detector parameters. A fill-factor of 48 % has been presented in [112] using the following techniques: the guard-ring width can be reduced, or even shared/overlapped with neighbouring SPADs, at the expense of increased interference between devices (commonly referred to as cross-talk); alternate rows of SPADs within the SiPM can be offset by half the pitch and brought into a *honeycomb* like structure, in doing so losing available space for accompanying quench circuitry; the use of the smaller passive rather than active quenching circuitry allow SPADs to be positioned closer together at the expense of extended dead-time; and finally it is possible to place passive quench circuitry outside the array at the expense of uncertainty in the quench timing. Furthermore, it is possible to avoid *N-well* spacing rules by using NMOS-only SPAD supporting circuitry at the expense of worsened timing jitter [113].

The SPAD dead-time issue has been significantly reduced by incorporating temporal compression into the SiPM architecture, implemented using a pulse-shortening monostable circuit at the output of each SPAD. This concept was developed simultaneously by [114] and as part of the research documented within this thesis [3][5]. The temporal compression ensures that the subsequent logic after the monostable circuit is only *locked* for the duration of the shortened pulse, even though the SPAD remains dead. The shortened pulse must still be long enough to propagate through the remaining output logic and successfully drive any subsequent circuitry such as counters or timers. This technique is only useful if there are sufficient elements available so that the probability of two photon events occurring within a SPAD dead-time but outwith the shortened pulse width are highly likely to fall on different detectors. In [114], pulse durations of around 1 ns are presented, which is over two orders of magnitude less than the SPAD dead-time of over 100 ns. This has the result of allowing photon rates to increase by over two orders of magnitude into the hundreds of megahertz region. Figure 2.21 shows an update of Figure 2.16, where a monostable circuit has been inserted at the output of each SPAD and the recombination is a logical OR of the outputs of these circuits to provide the SiPM functionality.

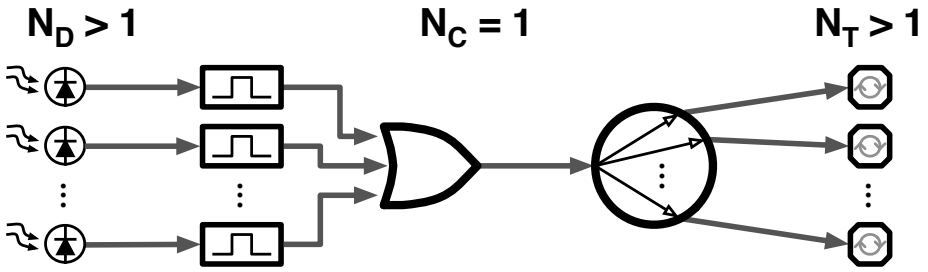


Figure 2.21: *Single multiple-element detector, multiple timing channel TCSPC architecture with monostable pulse-shortening at the output of each detection element.*

Even with the inclusion of the pulse-shortening monostable circuits, the OR-tree output of the SiPM still has a finite probability of being *locked* when two or more photon events arrive at two or more distinct detection elements within the duration of the shortened pulse-width. These photon event losses can be classified as a fourth form of TCSPC pile-up that is dependent not only on the pulse-width of the monostable outputs, but also on the number of detection elements within the SiPM. This additional form of pile-up will be discussed further in Chapter 3, where it will be investigated in detail.

[5]Further details on the implementation and performance of this circuit can be found in §4.3.3 and §5.3.4.

2.5.4 Time Digitisation

A thorough overview of time digitisation techniques and circuit implementations is provided in [66, 67, 115]. For use in a miniaturised CMOS TCSPC sensor that is capable of providing a photon throughput in excess of the current TCSPC pile-up limitation, the time digitiser must meet the following performance requirements: a low silicon area and minimal power consumption due to the requirement of multiple timers to overcome classical timer pile-up; a fast, ideally real-time conversion rate to overcome timer conversion pile-up; a time resolution in the picosecond to tens of picosecond region to allow correct resolvability of nanosecond fluorescent lifetimes; and finally a dynamic range of greater than 1 µs with good linearity to allow accurate resolvability of lifetimes up to hundreds of nanoseconds.

Full TAC-ADC conversion (introduced in Section 2.2.4) provides excellent time resolution, but is unsuitable due to the requirement of a high-speed 10-12 bit ADC, which would have a large area and long conversion time. Clocked delay lines (also introduced in Section 2.2.4) [116], Vernier delay lines (VDL) [117] and pulse shrinkers [118] can also be rejected due to their large silicon area and conversion times that both scale linearly with the dynamic range. Furthermore, passive interpolators [119] and time stretchers [120] – which provide excellent sub-gate delay time resolutions – also require a large silicon area, whilst the latter has an extended conversion time. This leaves three feasible options for the timer in this work, all using a *real-time* coarse-fine TDC approach [121]: the distributed clock architecture, which drives a coarse counter paired with either a small linear delay line [79, 122] or TAC [123] to provide the fine resolution; and the gated ring oscillator (GRO) approach, which uses a full cycle of the ring as the coarse resolution and its internal state as the fine resolution [124, 125]. However, distributing a high speed clock to all of the timers causes high static power consumption and is difficult to scale. Therefore the GRO approach, which is self-contained, has a low power consumption that is proportional to the dynamic activity and provides a *real-time* conversion is the preferred option for the given requirements.

The general architecture of a GRO based TDC is shown in Figure 2.22. The core of a ring oscillator is a linear delay line constructed using an odd number of inverters, whose final stage output is fed back to its input. The odd number of inverters provides inherent instability, so the circuit quickly settles into a state of oscillation caused by the finite gate delay through each element. A *START* pulse seeds the inverters with a known starting state and they are then frozen using a secondary *STOP* pulse. By counting the number of complete oscillations and sampling

the state of the inverter chain, it is possible to determine the time between *START* and *STOP* with the resolution being determined by the gate delay. For a TDC, the gate delay is typically designed to be the minimum of the process used for implementation. The number of complete oscillations is counted using a ripple counter and an encoder is used to convert the ring's fine state into a binary value. One of the advantages of this TDC architecture is that the area scales logarithmically (rather than linearly, as is the case for delay lines) with dynamic range by adding additional bits to the coarse counter. Although GRO-TDCs have a relatively low resolution compared to sub-minimum gate delay approaches (e.g. VDL), implementing them in advanced processes – which have tens of picosecond minimum gate delay – makes them suitable for fluorescence lifetime experimentation, as proven by the results from [10, 11, 96] which use $\approx$ 50 ps and $\approx$ 100 ps resolution GRO-TDCs.

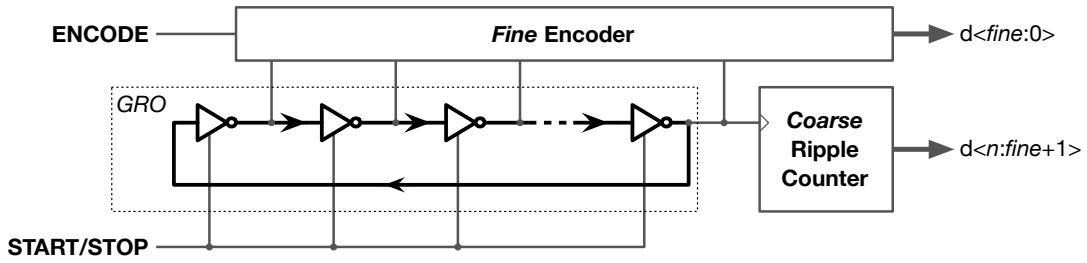


Figure 2.22: *Gated ring oscillator (GRO) time-to-digital converter (TDC) technique.*

As explained in Section 1.4, the $\approx$ 50 ps resolution, 50$\times$50 µm silicon area, real-time conversion GRO-TDC from [12, 22], plus the process it was developed in were made available for use within this research. Unlike standard ring oscillators described above, an even number of *differential* inverters is used in this implementation, as shown in Figure 2.23, where the NAND gates act as initialisation when R is low and as cross coupled inverters when R is high to ensure fully differential operation. Furthermore, S starts and stops the oscillation and T tri-states the outputs to prevent loading by the encoder. Oscillation is made possible by swapping the polarity of the feedback at the final element. This technique has the advantage of providing an exact binary power number of fine output states, so the four element differential inverter chain presented provides eight possible fine states that can be encoded to a 3-bit output. Despite having a real-time conversion, the TDC still requires a short period of time (tens of nanoseconds) to reset and prepare it for the next start-stop sequence, so timer dead-time is still an important consideration. A possible approach to overcoming this issue by using time-interleaved converters is introduced in the following section.

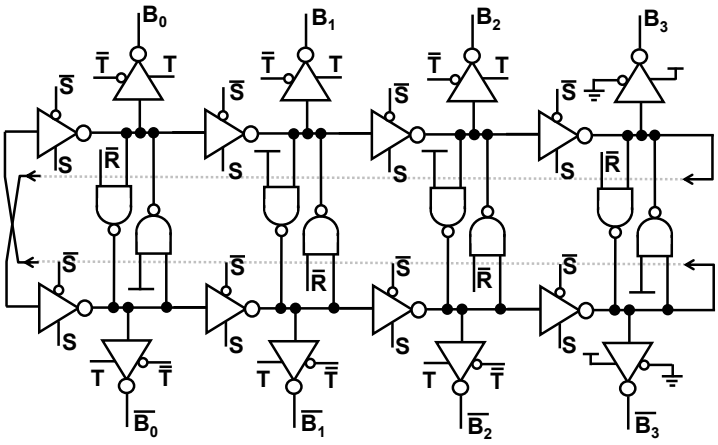


Figure 2.23: *Differential gated ring oscillator (GRO). [22]*

2.5.5 Time-Interleaved Converters

Reducing or removing timer conversion dead-time is critical to achieving TCSPC count rates in excess of the pile-up limit. Despite the TDC architecture and implementation introduced in the previous section having a dead-time much less than typical discrete TAC-ADC TCSPC timers, it is still advantageous to remove it completely. A technique exists for *analogue*-to-digital conversion where multiple converters (ADCs) are arranged in parallel to improve the speed and resolution of the system [126, 127]. Such an architecture is shown in Figure 2.24. Each converter operates at a frequency f_s/M – where f_s is the sampling frequency and M is the total number of converters – and their outputs are combined digitally at the output at a frequency f_s.

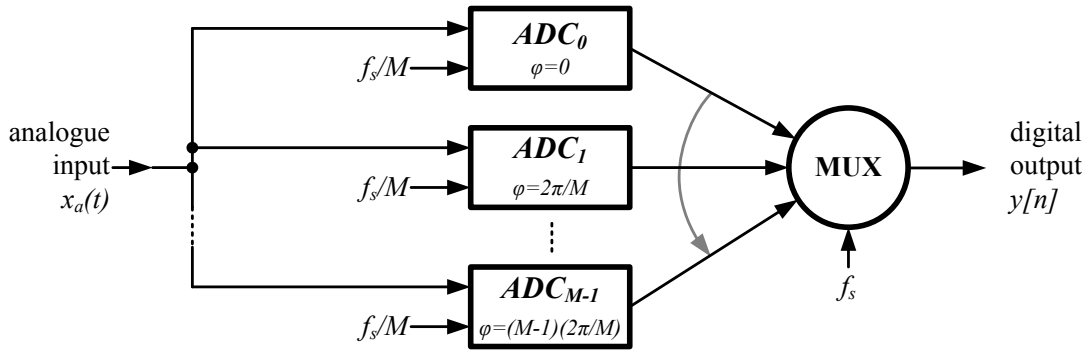


Figure 2.24: *Time-interleaved analogue-to-digital converter (TI-ADC) approach. [126]*

The technique allows the maximum frequency of the system to operate at M times the maximum frequency of a single converter by ensuring each ADC samples the input signal at a unique phase of f_s, given by $\varphi = (M-1) \cdot \frac{2\pi}{M}$. This has the effect of significantly increasing the bandwidth of the converter system proportional to M. The distribution of samples sequentially to different converters gives rise to the name of this technique: time-interleaved ADC (TI-ADC), and its operation is shown conceptually for $M = 2$ in Figure 2.25a. A time-interleaved TDC (TI-TDC) approach can therefore be used as a technique to remove timer dead-time, despite not operating on a continuous waveform ($a_x(t)$). This is shown in Figure 2.25b for $M = 2$, where f_s is equivalent to the excitation repetition rate and each TDC is capable of making a conversion and being reset at $f_s/2$, due to the dead-time of the timer (t_D). The number of timers required to ensure no apparent system dead-time is $M = \lceil \frac{t_D}{f_s} \rceil + 1$. Further simulation and analysis is necessary on the TDC architecture within its system to determine a value for t_D, and hence design for M given a maximum excitation frequency specification. This is presented in Section 4.4.2.

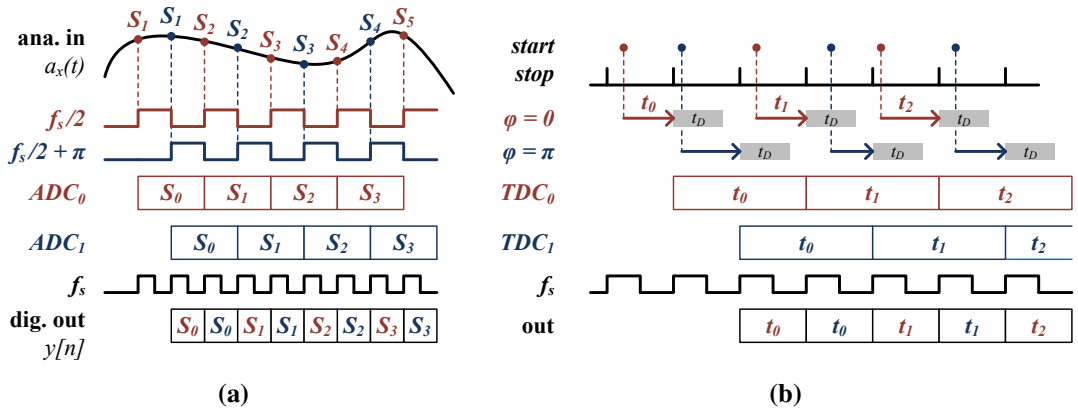


Figure 2.25: **(a)** *Time-interleaved ADC (TI-ADC) and* **(b)** *time-interleaved TDC (TI-TDC), each with* $M = 2$.

However, as shown in Figure 2.26, TI-ADCs suffer from three sources of distortion in addition to those already present in the individual converters themselves: phase, offset and gain mismatch. Phase mismatch is related to the correct distribution of clocks to ensure each converter samples the input at the correct point in time, however this is not an issue for a TI-TDC architecture, where we are not sampling a continuous waveform. Offset errors can be minimised for TI-TDCs by ensuring balanced distribution of the *start* and *stop* signals to the M timers. Finally gain error – which is the most serious source of distortion – can be

minimised in TI-TDCs by including some form of adjustable gain control at each TDC [12] or can be corrected for in the digital domain at the output of the TI-TDC [128]. The effects of timer gain mismatch on fluorescence lifetime will be studied in detail in Section 3.10.

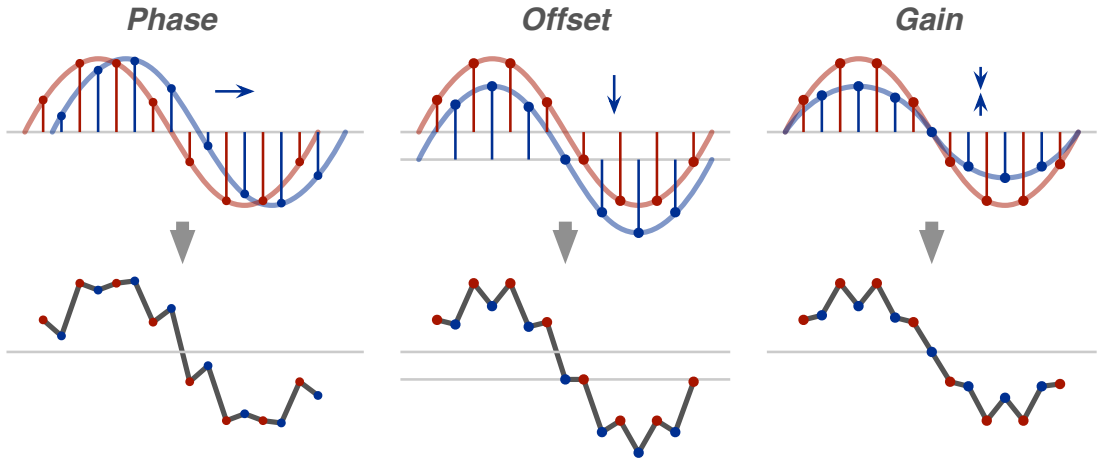


Figure 2.26: *Three sources of distortion from time-interleaved ADCs (TI-ADCs): phase, offset and gain mismatch.*

The TI-TDC approach can then be integrated into a multiple timing channel TCSPC architecture as shown in Figure 2.27. In this architecture, each *timing channel* consists of a TI-TDC sub-system with M converters which presents a zero dead-time *timer* to the event distribution, and so removes all conversion pile-up. The multiple timing channel architecture ($N_T > 1$) is then achieved by including T of these TI-TDC sub-systems to reduce the effect of classical timer pile-up. This architecture will introduce further possible gain errors between the multiple TDCs, which will form part of the investigation in Section 3.10.

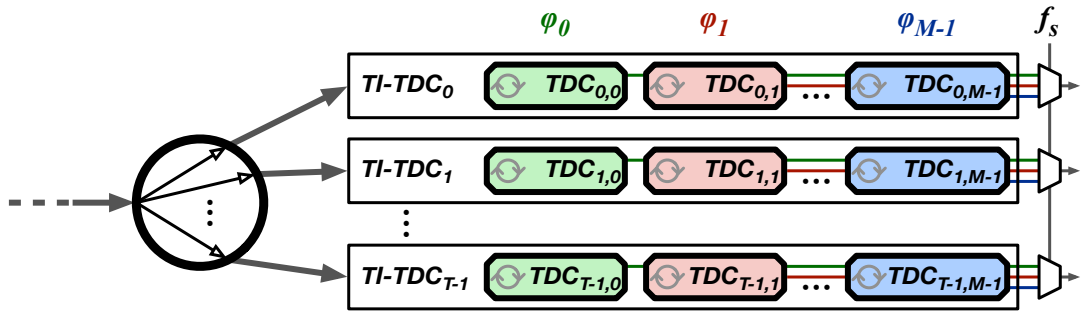


Figure 2.27: *Multiple TI-TDC timing channel architecture.*

2.6 Embedded Fluorescence Lifetime Calculation

2.6.1 Overview

Due to the high data rates expected when operating at increased photon throughputs greater than the current TCSPC pile-up limit, and the lack of bandwidth available at the chip interface in a miniaturised sensor, some form of TCSPC data compression is necessary. This can be achieved by performing a fluorescence lifetime calculation on-chip in *real-time*, which would further enable applications such as fluorescence activated cell sorting (FACS). The standard non-linear least-squares method [129] and maximum-likelihood estimation [130] algorithms used to perform curve fitting of captured TCSPC histogram data to known analytical decay equations require iterative floating point computation, more suitable to a generic microprocessor approach running a software algorithm [26, 27]. Therefore innovative non-iterative algorithms for processing raw TCSPC data to produce a fluorescence lifetime estimate are required. Ideally, these should easily be embedded on-chip using simple and compact digital design techniques, such as: addition, subtraction, multiplication and division by binary powers and compact lookup tables (LUTs). This section will describe three techniques available to achieve this [9] and discuss the advantages and disadvantages of each: rapid lifetime determination (RLD), which has been developed for time-gated fluorescence lifetime acquisition but can also be used on TCSPC data; integration for (lifetime) extraction method (IEM); and the centre of mass method (CMM). To finish, a short discussion on the precision performance of each of the above techniques is presented.

2.6.2 Rapid Lifetime Determination (RLD)

The most commonly used non-iterative fluorescence lifetime calculation technique is rapid lifetime determination (RLD) [25]. As introduced in Section 1.2.3, in its most basic form RLD counts the number of photon arrivals within two distinct time bins of equal width, between $0-t$ and $t-2t$. This basic form is referred to as two-gate RLD and is shown on the left of Figure 2.28. The fluorescence lifetime is calculated using Equation 2.9, where D_0 and D_1 are the total number of counts in each bin and Δt is the time-width of each bin. Clearly this calculation is not simple to perform with limited hardware, particularly due to the natural logarithm and the floating point division. However, data compression is achieved by accumulating events before periodically transmitting D_0 and D_1 to a host processor to complete the calculation.

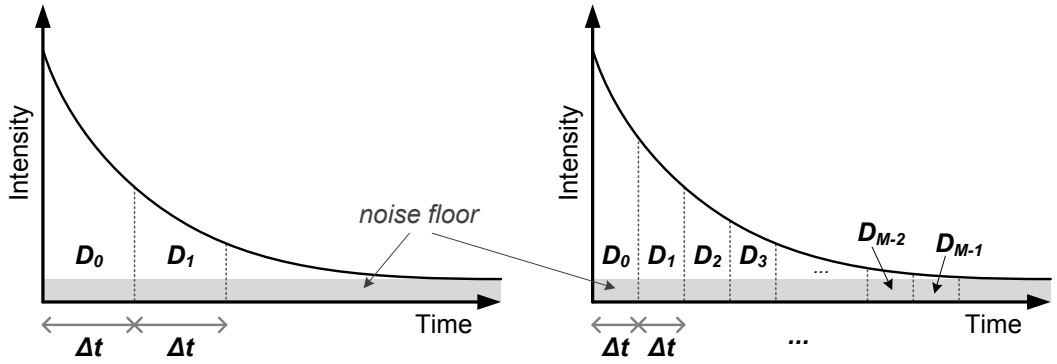


Figure 2.28: *Generalised two-gate rapid lifetime determination (RLD) (left) and multi-gate RLD or TCSPC acquisition (right).*

$$\tau_{\mathrm{RLD}}(t) = \frac{-\Delta t}{ln(D_1/D_0)} \tag{2.9}$$

Two-gate RLD can be expanded to *N*-gates with varying and even overlapping bin-widths, to improve the resolvability of the fluorescence decay [24]. For example, two-, four- and eight-bin approaches are suitable for single-, bi- and multi-exponential decays, respectively. This is achieved at the cost of increased calculation complexity, which typically reverts to iterative methods again but with a significantly reduced number of points compared to TCSPC. It has been suggested that *near-ideal* lifetime calculation is possible with optimally designed gating parameters [131], however due to the distribution of lifetimes in typical biological samples, this optimisation is non-trivial without prior knowledge of this distribution. Furthermore, as discussed in Section 1.2.3, increasing the number of time bins either increases the area required by the parallel counters, or decreases the acquisition speed if each time bin is captured sequentially. In both cases, increasing the number of time bins increases the data bandwidth requirements for an equivalent photon throughput.

For equal bin spacing and large N, the time-gated technique begins to approach that of TCSPC, where the time bin width (Δt) is equal to the TCSPC timing resolution, as shown by the graph on the right of Figure 2.28. For parallel acquisition, the high number of counters become a histogram memory, significantly increasing the data bandwidth. Furthermore, standard iterative post-processing techniques would be required to calculate the lifetime characteristic. However, non-iterative algorithms exist to perform fluorescence lifetime calculation in *real-time* using the raw TCSPC data and will be presented in the following section.

2.6.3 Integration for Extraction Method (IEM)

A new calculation method was proposed by Li *et al.* [132] as an alternative to RLD, called the integration for (lifetime) extraction method (IEM). The fluorescence lifetime is given by Equation 2.10, where h is the TCSPC resolution or time-bin width, N_j is the count number in the j^{th} time bin and $C = [1/3, 4/3, 2/3, ..., 4/3, 1/3]$ from Simpson's integration rule. The computation is less complex to perform in hardware than RLD, requiring an accumulator for the denominator and an up-down counter for the numerator. Furthermore, the division can be performed using a binary shift right when the denominator reaches a preset power of two. However, the circuit to perform the division by three, as required by the coefficients in C, adds unnecessary complexity.

$$\tau_{\text{IEM}}(t) = \left(\frac{\sum_{j=0}^{M-1} C_j \cdot N_j}{N_0 - N_{M-1}} \right) \cdot h \tag{2.10}$$

Fortunately, C can be replaced with $C' = [1/2, 1, ..., 1, 1/2]$ from Romberg's integration rule to provide a good estimate of the lifetime decay, where the divide by three circuit can be replaced with a selective binary right shift divide by two when required [96, 133]. Alternatively, using this simplification, the lifetime using IEM can then be given by Equation 2.11, where N_C is the total number of photon events. In this case, the hardware requirements are counters for N_C, N_0 and N_{M-1} plus an adder for $N_0 + N_{M-1}$ and a subtractor for $N_0 - N_{M-1}$. Again, the final division can be performed using a shift when the denominator reaches a power of two. The major advantage of IEM over RLD is much better immunity to uncorrelated background noise [96]. However, the position and width of the time bins must be carefully chosen to provide the most optimal calculation. In both cases, the value of h is fixed so is ignored from embedded calculation and the results are then provided as a function of h.

$$\tau_{\text{IEM}}(t) \approx \left(\frac{N_C - (N_0 + N_{M-1})/2}{N_0 - N_{M-1}} \right) \cdot h \tag{2.11}$$

2.6.4 Centre of Mass Method (CMM)

A second hardware efficient method of performing fluorescence lifetime calculation has been proposed by Li *et al.* [10, 134] using a centre of mass calculation termed the centre of mass method (CMM). For a single exponential decay, the lifetime can be calculated by the average arrival time of all detected photons and can be described in the continuous and discrete time domains by Equation 2.12. It is possible to use the discrete version of this equation on captured TCSPC histogram data. The hardware requirements for CMM are an accumulator to add together arriving TCSPC time-stamps ($j \cdot N_j$) and a counter to count the total number of events (N_C). As with IEM, the division can be performed using a binary right shift when a power of two events have been counted. The $\frac{1}{2}$ can be ignored for simplicity. In practice the peak of the fluorescence decay does not lie exactly at time $t = 0$, so it is necessary to introduce a windowing function that only includes TCSPC events in the calculation if they lie between preset limits *FIRST* (F) and *LAST* (L). In a reversed start-stop system, *LAST* defines the position of the decay peak and the result must be subtracted from it to retrieve the final lifetime calculation.

$$\tau_{\text{CMM}}(t) = \frac{\int_0^\infty t \cdot f(t)\,\mathrm{d}t}{\int_0^\infty f(t)\,\mathrm{d}t} = \left(\frac{\sum_{j=0}^{M-1} j \cdot N_j}{N_C} + \frac{1}{2} \right) \cdot h \tag{2.12}$$

Unlike IEM, in its basic form the result of Equation 2.12 is very dependent on any uncorrelated background noise. The windowing function described above can minimise this effect by rejecting any noise that lies outside the limits, however correction of the final result is still required. The contribution of the noise to the calculation is completely uncorrelated so can be assumed to have a uniform distribution, as shown by the grey area in Figure 2.28. The background corrected CMM lifetime is therefore calculated using Equation 2.13, where M is the total number of time bins ($L - F + 1$) and N_b is the noise contribution of each. As expected, this correction requires good calibration of the expected background contribution (N_b). The static values for $(F + L) \cdot N_b \cdot M/2$ and $M \cdot N_b$ can be provided, so only two additional

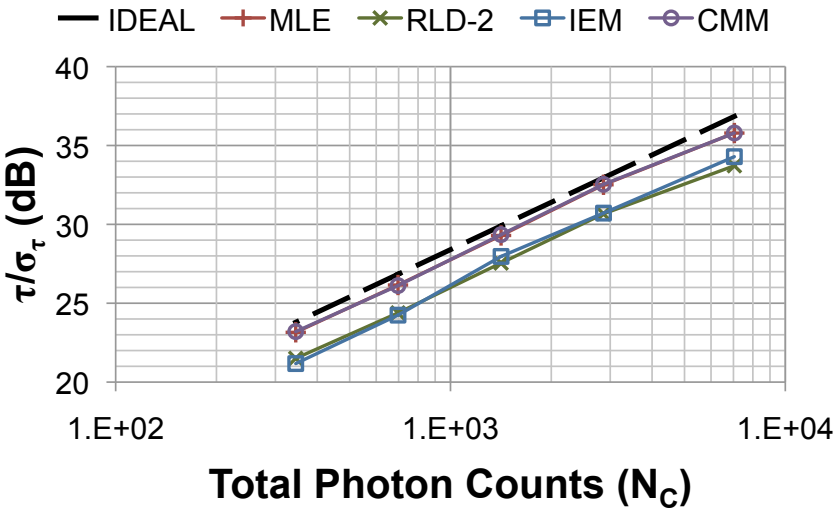
subtractors are required to implement this correction in hardware. CMM is a *near-ideal* single exponential lifetime estimator, assuming correct background calibration, and is 100 % photon efficient as every *detected* photon event arrival contributes to the final calculation, making it an ideal choice for high throughput applications.

$$\tau_{\text{CMM\_BC}}(t) = \left(\frac{\sum_{j=0}^{M-1} j \cdot N_j - (F+L) \cdot N_b \cdot M/2}{N_C - M \cdot N_b} \right) \cdot h \tag{2.13}$$

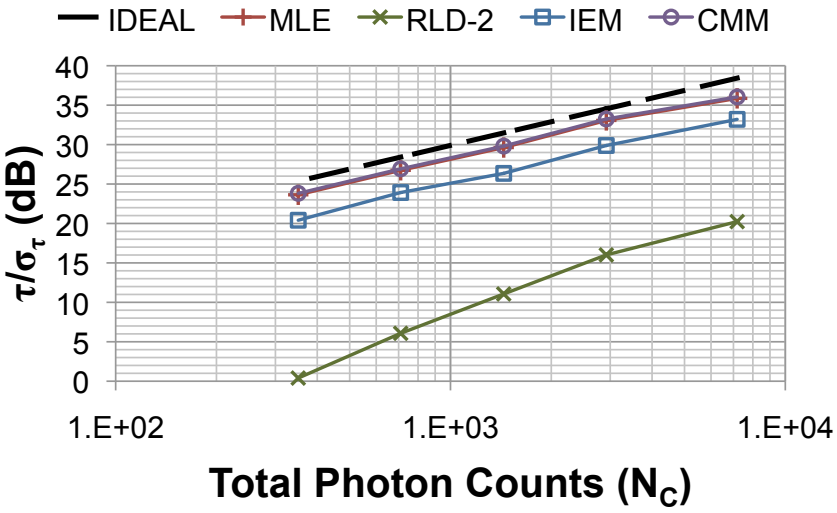
2.6.5 Calculation Precision

In addition to keeping the hardware requirements for the calculation low, it is also important to base the selection of an embedded algorithm on its performance. This section discusses the relative precision available using maximum-likelihood estimation (MLE), RLD-2, IEM and CMM on a sample data set. The data set is real lifetime data of a $\approx$ 1.8 ns fluorophore (Rhodamine B) captured using an integrated SPAD/TDC device [12], which will provide a good basis for comparison to the system being developed as part of this work.

Figure 2.29 shows the results of these precision tests for all calculation techniques and with different measurement windows of 4.1τ (Figure 2.29a) and 17τ (Figure 2.29b) [9]. The ideal shot-noise limited precision is shown for comparison by the dashed line. The window of 4.1τ is chosen as the optimal window for RLD-2, where it is shown to have comparable precision performance to IEM. However, changing the measurement window to 17τ has a considerable affect on the precision of RLD-2, which further highlights its sensitivity to the lifetime being measured. Expansions to the RLD calculation, including additional, unequal and/or overlapping bins show improved precision performance. However, this is achieved at the cost of increased hardware requirements and complexity, whilst never achieving better precision than least-squares or maximum-likelihood. IEM shows consistent precision performance, better than or equal to that of RLD-2. However, CMM is clearly the most precise non-iterative technique presented, equalling that of the iterative maximum-likelihood curve fitting approach for a single exponential decay over a range of measurement windows.



(a)



(b)

Figure 2.29: *Precision plots of a TCSPC captured Rhodamine B data set comparing MLE, RLD-2, IEM and CMM lifetime calculation techniques with measurement windows of* **(a)** *4.1τ and* **(b)** *17τ. [9]*

2.7 Conclusions

This chapter has provided a detailed review of single-channel TCSPC, its limitations in terms of pile-up and techniques to overcome and/or minimise these limitations using hardware and software approaches. It continued by looking at variations of multi-module TCSPC approaches with a view to embedding one of the system architectures on an integrated CMOS sensor. Finally, the technologies, structures and algorithms available to perform efficient single photon detection, picosecond resolution timing and fluorescence lifetime calculation or data compression on standard CMOS are reviewed.

Techniques have been proposed to overcome the three primary pile-up limitations of TCSPC. Firstly, *classical* timer pile-up can be minimised by adding multiple timing elements to time more than one photon event in a given excitation period. Secondly, non-extending timer dead-time (conversion) pile-up can be removed by using a time-interleaved TDC (TI-TDC) approach, which operates in a similar way to the more common TI-ADC architecture. Finally, extending detector dead-time pile-up can be minimised by incorporating temporal compression in a digital silicon photomultiplier (SiPM) architecture to reduce the apparent dead-time of the individual SPAD elements by two orders of magnitude. The combination of these individual techniques and approaches leads to the architecture shown in Figure 2.21, where the SiPM architecture ($N_C = 1$) is chosen to maximise the fill-factor of the single photon sensitive area of the device. Although these techniques provide the possibility of overcoming TCSPC pile-up, care must be taken with the multiple channel TI-TDC architecture to minimise or correct for mismatch. Furthermore, the finite compressed pulse-width of the SiPM output adds additional complexity to detector dead-time pile-up and will still limit the available photon throughput.

The availability of the SPAD structure, core TDC circuit and the process that each was developed in allows the research presented in this thesis to focus solely on the architecture and practical applications that are suited to it. The techniques to increase fill-factor in SiPM architectures, as introduced in Section 2.5.3, are an area of current cutting-edge research. Therefore to minimise the risks associated with these unproven techniques, a safe SiPM fill-factor will be designed, ensuring that the architecture can be fully tested. Furthermore, the gated ring oscillator core of the TDC will be used *as-is* with modifications only being necessary to its peripheral circuits by adding bits to the coarse ripple counter to extend dynamic range and to its interfacing logic to integrate it into the proposed architecture.

Due to the increase in data bandwidth required by a TCSPC sensor operating beyond the pile-up limit, and the lack of transmission lines at the interface of a miniaturised implementation, some form of fluorescence lifetime calculation or data compression is deemed necessary. Although not accurate in fully describing real biological behaviours, single-exponential decay models are useful to contrast different types of fluorophores. For diagnostic applications, obtaining high-speed lifetime contrast can be more important than determining the absolute values of lifetimes [135]. Therefore, the 100 % photon efficient, high-throughput CMM calculation – which provides single exponential precision performance comparable to the commonly used iterative least-squares and maximum-likelihood methods – is clearly the most suitable given this simplification.

The following chapter will briefly redefine the sources of pile-up arising from the chosen architecture before performing an in-depth study of the system and calculation technique using both mathematical analysis and modelled simulations. It will provide validation that the architecture is in fact capable of achieving an increase in throughput, define a parameter set for the number of detectors in the SiPM, compressed pulse-width and number of timers, and finally it will present the expected performance capability of this chosen parameter set. In addition to this, it will also investigate the effect of mismatch errors in the multiple channel TI-TDC architecture.

3
PILE-UP IN AN INTEGRATED TIME-CORRELATED SINGLE PHOTON COUNTING ARCHITECTURE

3.1 Introduction

Chapter 2 introduced a number of architectures for increasing photon throughput in TCSPC fluorescence lifetime sensing using different configurations of detectors, timers and signal processing. The architecture chosen to be implemented is a multiple element detector arranged as a digital SiPM with a single pulse-shortened output, combined with multiple time-interleaved TDC (TI-TDC) timing channels and a centre-of-mass method (CMM) algorithm to process photon events in parallel. A thorough investigation of this system architecture using modelling, simulation and theoretical analysis is now described to help understand its intricacies and to provide evidence that the chosen design is in fact capable of increasing photon throughput in fluorescence lifetime experimentation.

The chapter will begin with a detailed description of how the system architecture is modelled, before using it to simulate the classic TCSPC setup, with a single detector and single timer. This will highlight the current limitations of such a system and help to define an acceptable error in accuracy for the lifetime calculation. The different architecture design choices: number of timing channels, shortened SiPM output pulse-width, number of detectors and detector dead-time will then be varied in different experimental configurations to study the effect each has on the resulting TCSPC histograms and lifetime calculations. In particular, these simulations, together with theoretical analysis, will look at how and when photons are lost due to the different forms of pile-up. Proposals are then made for the controllable factors of the design to allow throughput to exceed the excitation repetition rate, providing a specification for the implementation of an integrated sensor. Finally, precision of the system is studied before the effects of timing mismatch in the proposed multiple-TDC architecture are investigated.

3.2 Modelling Pile-up in TCSPC

3.2.1 Overview

A system model is required to verify the architecture's functionality and to help understand the effects of the available design choices. This section begins by introducing the sources of pile-up in the chosen architecture, followed by a description of the architectural and experimental variables of such a system before discussing the assumptions and simplifications that can be made. The implementation of a transient MATLAB system model, that focuses on the many aspects of pile-up losses and their effect on TCSPC data and CMM lifetime calculation, is then described. Finally, an investigation strategy is outlined which forms the structure of the following sections in this chapter.

3.2.2 Sources of Pile-Up

The sources of TCSPC pile-up in the chosen architecture from Chapter 2 are shown conceptually in Figure 3.1. Although similar, the different forms of pile-up resulting from this architecture do not match those typically found in single channel TCSPC, as introduced in Section 2.3. Therefore the model will explicitly measure the amount of each form of pile-up as defined in this figure, which are described below.

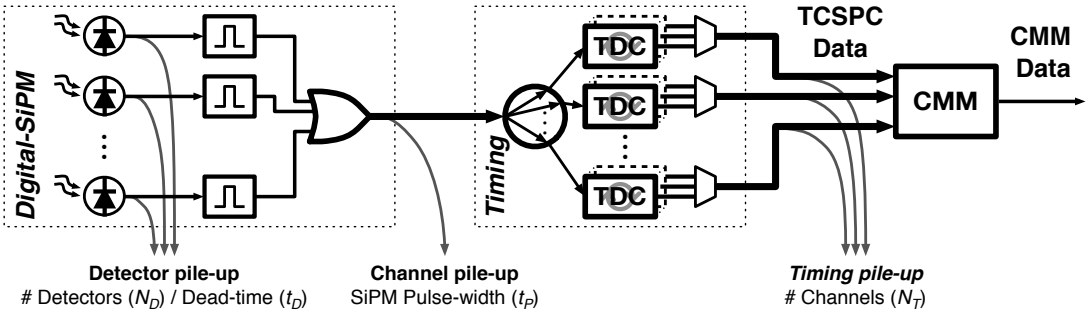


Figure 3.1: *Sources of pile-up in the chosen system architecture.*

The effect of classical timer pile-up is minimised with the introduction of N_T TI-TDC timing channels, however timer pile-up will still occur if there are more photons in an excitation period than there are timing channels available to process them. Timer dead-time (t_D), or conversion time, has been removed completely by the use of TI-TDC timing channels, assuming $M \geq \lceil \frac{t_D}{f_s} \rceil + 1$ (see Section 2.5.5), therefore will not be included in the model.

Despite detector dead-time (t_D) pile-up being minimised by the introduction of multiple detection elements (N_D), over which the emission photons are evenly distributed, it will still be a source of pile-up. Finally, as described in Section 2.5.3, a fourth form of pile-up is introduced in the routing channel between the SiPM and the timers. This is caused by pulse overlap within the OR-tree recombination of the SiPM and is dependent on the shortened pulse-width (t_P) of each detected photon event.

The pile-up that occurs at the detector and at the output of the SiPM are very similar, but on different time-scales, as $t_D \gg t_P$. If a second event occurs at the same detector within the dead-time, t_D, of a previous event, then the second event will be lost. Similarly if a second event occurs at *any* detector within the pulse-width, t_P, of a previous event, then the second event will be lost. As the photon-rate is increased, the probability of these pulse overlap conditions increases. For a fixed total photon-rate, the amount of detector pile-up is reduced by increasing the number of detectors, N_D, so long as light is distributed equally between these detectors. For a fixed number of detectors however, the amount of both forms of pile-up will be photon-rate dependent and will be investigated in the following sections of this chapter.

To minimise pile-up losses, the architecture should ideally have short detector dead-times (t_D) and shortened pulse-widths (t_P) combined with a large number of detectors (N_D) and timing channels (N_T). However, process and implementation constraints will limit how short the detector dead-times and shortened SiPM output pulse-widths can be. Furthermore, increasing the number of detectors and timing channels will have practical area constraints. Each TDC will be $\approx$ 2,500 μm^2 [12], whilst the SPAD and its accompanying circuitry will be $\approx$ 500 μm^2 (based on layout trials).

3.2.3 Parameters

All of the variables relating to the design of the proposed integrated sensor architecture, including those highlighted in Figure 3.1 and described above, are shown in Table 3.1. In addition to the parameters already introduced (N_D, t_D, t_P and N_T) there is a dark count rate (*DCR*) or uncorrelated noise contribution to the detected optical signal and the timing channels have a base resolution (R_T), operate either in a standard or reversed start-stop mode (rev) (see Section 2.2.2) and contain N_M TDCs to achieve TI-TDC operation (see Section 2.5.5). Furthermore, to model timing mismatch, the TDC resolutions (R_T) will be modelled using a Gaussian distribution with a fixed mean but variable standard deviation (σ_T).

Variable	Description	Simulation Value(s)
N_D	Number of SPAD detectors in the SiPM	1 – 1,024
t_D	SPAD dead-time	10 – 50 ns
DCR	SiPM Dark Count Rate (DCR)	0 – 1 kHz
t_P	Shortened SiPM output pulse-width	$<$ 1 ns
N_T	Number of timing channels per excitation	1 – 100
N_M	Number of TDCs per TI-TDC timing channel	2
R_T	Resolution of timing-channels	50 – 400 ps
σ_T	Standard deviation of TDC resolutions	1 %
rev	Reverse start-stop (Boolean)	true / false

Table 3.1: *Device specific parameters of the pile-up model.*

As well as modelling the integrated system architecture, it is necessary to also include the experimental variables that have an impact on the outcome of TCSPC or fluorescence lifetime results. All of the variables relating to the configuration of the experimental conditions in this investigation – including the lifetime under observation (τ) and the average photon rate (μ) as a fraction of the excitation rate of the laser (f_E) – are detailed in Table 3.2.

Variable	Description	Simulation Value(s)
τ	Lifetime Value	1 – 100 ns
μ	Photon rate (fraction of excitation rate)	0.01 – 100.0
C_P	Ideal peak histogram counts	100 – 10,000
f_E	Excitation Frequency	5 – 20 MHz
T_P	Excitation (peak histogram) position	0 – $1/f_E$

Table 3.2: *Experiment specific parameters of the pile-up model.*

Lifetime values (τ) will always be kept below one fifth of the excitation period ($\tau < 0.2/f_E$), to both allow the lifetime to be accurately resolved and to keep cyclic effects (events in one excitation affecting those in the next) to a minimum. The number of photons occurring within a single excitation period follows a Poisson distribution and is defined by the average value, μ. Unrealistically high values of μ (up to 100) are used to help fully understand the architecture under extreme conditions. The offset between optical excitation and electrical synchronisation defines the position of the peak in the histogram, which can be varied between 0 and the period of the excitation rate. This parameter will only be changed when the system is configured in a reversed start-stop mode (rev), otherwise it will be set to 0, for consistency.

3.2.4 Implementation

A number of simplifications and assumptions are made to reduce the parameter set of the system model. This will keep simulation times low and focus the investigation on the aspects of the architecture that are most important for understanding high photon throughput fluorescence lifetime sensing and reduction of the various forms of pile-up.

- In practice, the dark count rate (*DCR*) of each individual SPAD will follow a distribution similar to that shown in Figure A.1. However, for the purposes of simplifying the model, it is assumed to be a single aggregate value, equal to the sum total *DCR* of all SPADs. This single value will be uniformly distributed between all SPADs in any given configuration.
- Variation of detector dead-times (t_D) and shortened pulse-widths (t_P) is negligible, so is assumed to be zero. As well as simplifying the model by removing extraneous parameters, this provides a better match to the mathematical theory introduced later in the chapter, which assumes a fixed value for t_P.
- Timing offsets are assumed to be ideal as they can be made negligible by efficient design and layout to ensure balanced distribution of the start and stop signals.
- The instrument response, which includes detector and timing jitter, is also assumed to be ideal, primarily as it will not affect pile-up performance.

The model is implemented as a function in MATLAB, the source code of which is provided in full in Appendix A.2. Figure 3.2 shows a simplified outline of the major steps in the function, highlighting where each of the main input parameters are used. Simulation scripts are used to run the model using different parameters, depending on the current focus of investigation.

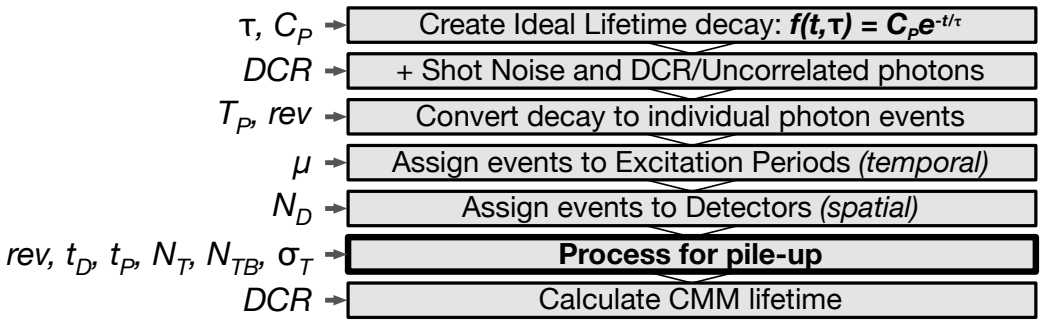


Figure 3.2: *Outline of the major steps of the* MATLAB *pile-up model.*

The function begins by creating a decay histogram at the simulation's base resolution using the provided lifetime (τ) and peak-count (C_P) parameters (Appendix A.2:lines 41-50). Next, photon shot noise and uncorrelated noise (*DCR*) is added to this ideal decay histogram, requiring the number of excitation periods and hence total time of the experiment to be calculated (A.2:53-74). Although *DCR* can be set to 0, shot noise will always be included in the decay histogram as it will always exist in practical experiments.

The high resolution decay histogram is then converted into an array of discrete individual photon event (micro) time-stamps (A.2:82-105). Each of these individual events is assigned a random excitation period (macro time-stamp) between 1 and the total number of excitation periods, defined by μ, ensuring events are distributed temporally according to a Poisson distribution. The array of pairs of micro and macro time-stamps are then ordered temporally by their macro time-stamp (A.2:108-122). Next, each individual event is distributed spatially by being assigned a uniformly random detector, between 1 and N_D (A.2:123-131). Each event now has both micro and macro time-stamps as well as an assigned detector.

The model then processes each event in turn, as per the algorithm shown in Figure 3.3 and as implemented on lines 164-286 in Appendix A.2. The algorithm begins by checking if the current event is still in the same excitation period as the previous event (macro time-stamp), before testing for each form of pile-up (3.3:19-22). If pile-up has *not* occurred for a given event, its micro time-stamp is converted from the simulation's base resolution to the resolution of the current timing channel, using R_T, σ_T and an event counter (j) before being added to a new histogram of *processed* events (3.3:24-26). If pile-up has occurred, the event is not added to the new histogram and a counter recording the type of pile-up is incremented (3.3:27-29). The order in which pile-up is checked is important, as it must follow the same order as shown in Figure 3.1, where detector losses come first, then channel losses and finally timing losses. When a new excitation period is reached, the event counter is reset and the process is repeated (3.3:31-39).

Finally, the fluorescence lifetime of the data in the recorded histogram is calculated using the centre-of mass method (CMM) with background correction (see Section 2.6) using the known *DCR* (A.2:328-353). As well as the model returning the *processed* histogram, background corrected CMM result and statistics on each form of pile-up, the contribution that each timing channel makes to the total decay histogram is recorded. This provides an insight into how many timing channels are required for a given set of parameters to help with the design choice of N_T.

```
% events(1,:) = Photon micro-times
% events(2,:) = Photon macro-times
% events(3,:) = SPAD event originated from
% R_T2 = Individual timing channel resolutions
% decay2_PU_ch = Contribution of each timing channel to decay
% decay2_PU = Final decay histogram
% PU_D = detector dead-time pile-up
% PU_P = pulse-width pile-up
% PU_T = timing pile-up
% j = inter-excitaiton period event counter

neg = rev ? -1 : 1;

prv = [ -t_P, 0 ];
prv_D = zeros(1,N_D) - t_D;

for i=1:N_events
  cur = laser_period*events(2,i) + neg*events(1,i);
  if ( prv(2) == events(2,i) )
    if ( cur - prv_D(events(3,i)) ) > t_D
      if ( cur - prv(1) ) > t_P
        if ( j < N_T )
          j = j + 1;
          ch = j + N_T*mod( events(2,i), N_M );
          bin = ceil( events(1,i) / R_T2( ch ) );
          decay2_PU_ch(ch, bin) = decay2_PU_ch(ch, bin) + 1;
        else PU_T = PU_T + 1;
      else PU_P = PU_P + 1;
    else PU_D = PU_D + 1;
  else
    j = 0;
    if ( cur - prv_D(events(3,i)) ) > t_D
      if ( cur - prv(1) ) > t_P
        j = j + 1;
        ch = j + N_T*mod( events(2,i), N_M );
        bin = ceil( events(1,i) / R_T2( ch ) );
        decay2_PU_ch(ch, bin) = decay2_PU_ch(ch, bin) + 1;
      else PU_P = PU_P + 1;
    else PU_D = PU_D + 1;

  prv = [ cur, events(2,i) ];
  prv_D(events(3,i)) = cur;

decay2_PU = sum(decay2_PU_ch);
```

Figure 3.3: *Pseudo-code detail of pile-up processing in the* MATLAB *system model. For further details, see Appendix A.2, lines 164-286.*

3.2.5 Investigation Strategy

The aim of this investigation is to verify operation of the proposed architecture and to define a specification of the parameters for implementation of the integrated fluorescence lifetime sensor. It begins by using the model to look at a typical TCSPC setup with a single timing channel ($N_T = 1$) and no processing dead-time. This will provide information on the acceptable error in accuracy of the lifetime calculation to guide the remainder of the chapter. Each of the three primary aspects of the architecture are then studied individually, beginning with the number of timing channels (N_T), then the pulse-width (t_P) and finally the detector's parameters (N_D, t_D and *DCR*). When looking at each of these aspects, the parameters not being studied are idealised ($N_T, N_D = \infty$ and $t_P, t_D, DCR = 0$). Finally, the effects of all parameters combined will be investigated before proposals are made for the final specification.

Each of these investigations (timing-channels, pulse-width, detectors and combined effects) will follow a similar strategy. To begin with, representative histograms are captured for fixed values of μ, to highlight how each form of pile-up distorts the decay in different ways. Then CMM calculations are performed while sweeping μ to show how pile-up affects the accuracy of lifetime calculation as the photon-rate is increased beyond typical values. In addition to calculating lifetime values whilst sweeping μ, the three forms of pile-up are also captured to monitor how they are affected by the architectural choices. Finally, the available photon throughput rate (μ), for assumed acceptable lifetime calculation errors[1] of 1 % and 5 % (chosen based on results from Section 3.3.3), is plotted against the parameter currently under investigation. Theory is introduced when possible to describe the expected histogram and CMM results, helping to validate the model.

The absolute value of the lifetime is irrelevant, but its relative value in terms of R_T, f_E, t_D and t_P is important. For the purposes of consistency in these studies, $R_T = \tau/100$ and $\tau = 0.1/f_E$, whilst t_D and t_P are varied. Furthermore, for the initial investigations $C_P = 10,000$ to minimise statistical errors. Simulation scripts are written to vary the parameters for each of the proposed investigations, as described above. The MATLAB random number generator (RNG) is reset before each experiment to provide consistent and repeatable results. Furthermore, results presented will normalise the lifetime to an ideal factor of 1.0, by dividing the result of CMM from the histogram suffering from pile-up by the result of CMM calculated from a histogram with no pile-up (ideal parameters) (see A.2:128-145 and 281-321).

[1] Unless otherwise stated, *error* refers to an error in *accuracy* of the lifetime calculation.

3.3 Single Timing-channel TCSPC

3.3.1 Overview

To begin with, the model is used to investigate a classic single timing channel TCSPC architecture ($N_T = 1$), but with negligible timing dead-time (which is not included in the model). Furthermore, the detector configuration is assumed to be ideal ($N_D = \infty$ and $t_P, t_D, DCR = 0$). This represents an optimal configuration of a single-channel TCSPC system, where *all* pile-up is caused by the inability of the hardware to process more than the first event in each excitation period (*classical* TCSPC pile-up).

3.3.2 TCSPC Decay Histograms

The graph in Figure 3.4 shows the effect of increasing μ from 0.01 to 10.0 on the resulting captured TCSPC decay histogram using both the model (solid) and the theory introduced in Section 2.3, described by Equation 2.1 (dashed). To ease comparison, the values in the histograms are normalised by C_P to provide a value of ≈ 1.0 in the peak bin. It is important to note that it is much faster to acquire data for high values of μ as the time taken to acquire each set of histograms is inversely proportional to μ.

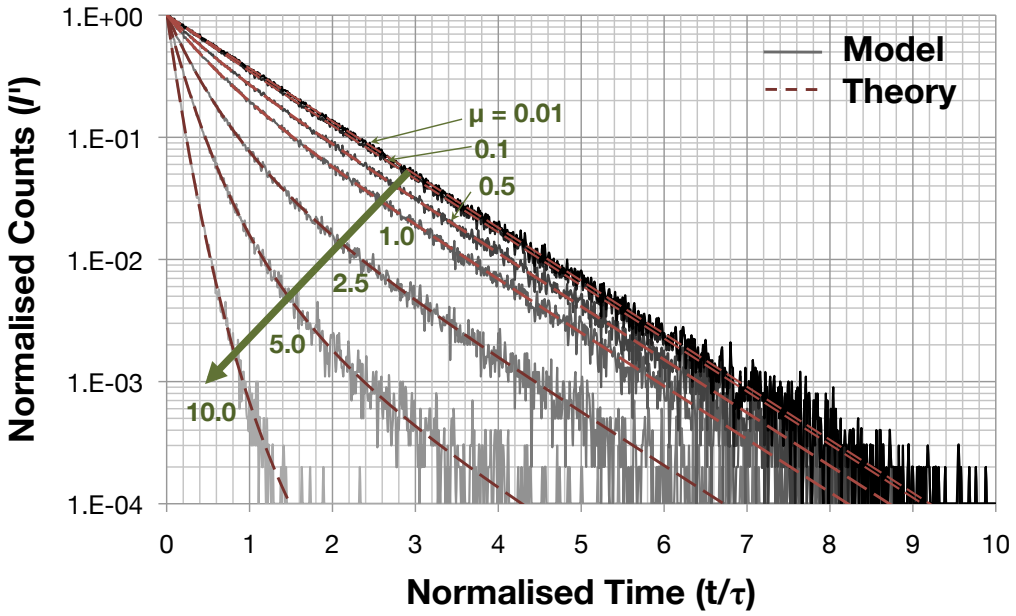


Figure 3.4: *Effect of increasing μ on the captured histogram for single timing channel TCSPC using the model (solid) and theory (dashed).*

This result matches what is expected from the theoretical description of classic, single timing channel TCSPC, as introduced in Section 2.3, where the histogram is distorted towards shorter times as μ is increased. A worst case error of (+/- 3%) is found between theory and simulation, caused by quantisation and the inclusion of photon shot-noise in the model.

3.3.3 CMM Calculation

The effect that this form of pile-up has on the fluorescence lifetime calculation using CMM can be shown theoretically. By combining Equations 2.1 and 2.12, the fluorescence lifetime by CMM is given by Equation 3.1[2], where T is the width of the timing window, which in this case is 10τ, and $\mathrm{Ei}(x)$ is the exponential integral, described by $\mathrm{Ei}(x) = \int_{-\infty}^{x} e^t/t \ \mathrm{d}t$. The portion of photons lost to classical TCSPC pile-up is introduced in Section 2.3 (see Equation 2.6).

$$\tau'_{\mathrm{CMM}}(\mu) = \frac{e^{-\mu}\left[\tau \cdot \mathrm{Ei}\left(\mu\right) - \tau \cdot \mathrm{Ei}\left(\mu \cdot e^{-T/\tau}\right) - T \cdot e^{\mu \cdot e^{-T/\tau}}\right]}{1 - e^{-\mu\left(1-e^{-T/\tau}\right)}} \tag{3.1}$$

The graph in Figure 3.5 shows the effect of increasing μ on the accuracy of the calculated lifetime (left) and on the number of photons lost to pile-up (right), using both the model ($\times$/$\circ$) and theory (solid/dashed). As expected, the model and theory are shown to match. Performing the more commonly used MLE calculation on the same dataset produces identical accuracy results. In addition to the equivalent precision results (see Figure 2.29 in Section 2.6.5), this further reinforces the suitability of CMM to calculate single exponential lifetimes.

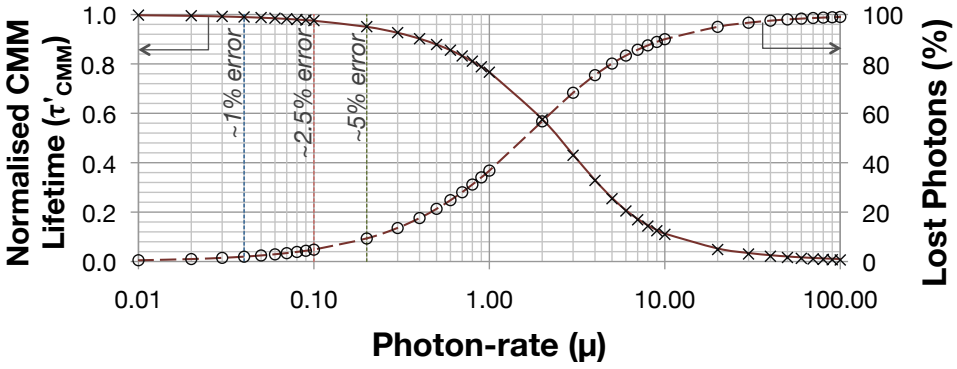


Figure 3.5: *Effect of increasing μ on lifetime calculation (left) and photon loss (right) for model ($\times$ and $\circ$) and theory (solid and dashed) in single timing channel TCSPC.*

[2] A full derivation of Equation 3.1 is provided in Appendix A.3

At a photon rate of $\mu = 0.10$, which is typically given as the maximum throughput possible for TCSPC without major distortion due to pile-up [18, 19], a CMM lifetime error of 2.5 % is calculated for a loss of just under 5 % of photon events. The remaining sections in this chapter will present the maximum available throughput (μ_{max}) possible for acceptable calculation errors of 1 % and 5 %, which lie on either side of this nominal 2.5 % error. In the single channel case, maximum throughputs (μ_{max}) of 0.04 and 0.20 are achievable for these errors of 1 % and 5 %, respectively. It should be noted that the portion of photons lost due to pile-up will be worse with the inclusion of conversion dead-time in the theory and model, so these results are the best case scenario assuming a single timing channel. In this case, at a photon-rate equal to the excitation rate ($\mu = 1.0$), the CMM lifetime calculation has an error of over 23 % and more than 37 % of photon events are lost to pile-up.

3.4 Timing Channel Pile-Up

3.4.1 Overview

The first parameter in the integrated fluorescence lifetime sensor to be investigated is the number of timing channels available per excitation period (N_T). As with Section 3.3, the detector configuration is assumed to be ideal ($N_D = \infty$ and $t_P, t_D, DCR = 0$). Although this configuration is not feasible practically, it allows the effect of N_T to be investigated without interference from other aspects of the architecture.

3.4.2 TCSPC Decay Histograms

In this idealised case, it can be assumed that within each excitation period, the mth photon event arrival is processed by the mth timing channel. This is not the case when $t_P, t_D > 0$, as events lost to pulse-width or dead-time overlap cannot be processed by any timing channel. The contribution that each photon event, and hence each timing channel, m, makes to the total decay histogram is given by Equation 3.2 [1] (see Appendix A.4, assuming t_P = 0), which is equal to Equation 2.1 for $m = 1$ and becomes recursive for $m > 1$.

$$I_T(\mu; t; m) = \begin{cases} e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)} & m = 1 \\ \frac{\mu}{m-1} \cdot \left(1 - e^{-t/\tau}\right) \cdot I_T(\mu; t; m-1) & m > 1 \end{cases} \tag{3.2}$$

The graph in Figure 3.6 shows the effect of varying the number of timing channels available per excitation period on the resulting TCSPC histograms for an excessively high photon-rate (μ) of 10.0, using both the model (solid) and theory (dashed). The theoretical curves are created by summing I_T for $m = 1 \rightarrow N_T$ for each N_T. As with Section 3.3, the only discrepancies between model and theory are quantisation and the inclusion of photon shot-noise in the model. In this particular instance ($\mu = 10.0$), the decay histogram is shown to recover to an approximately ideal decay for $N_T \geq 16$.

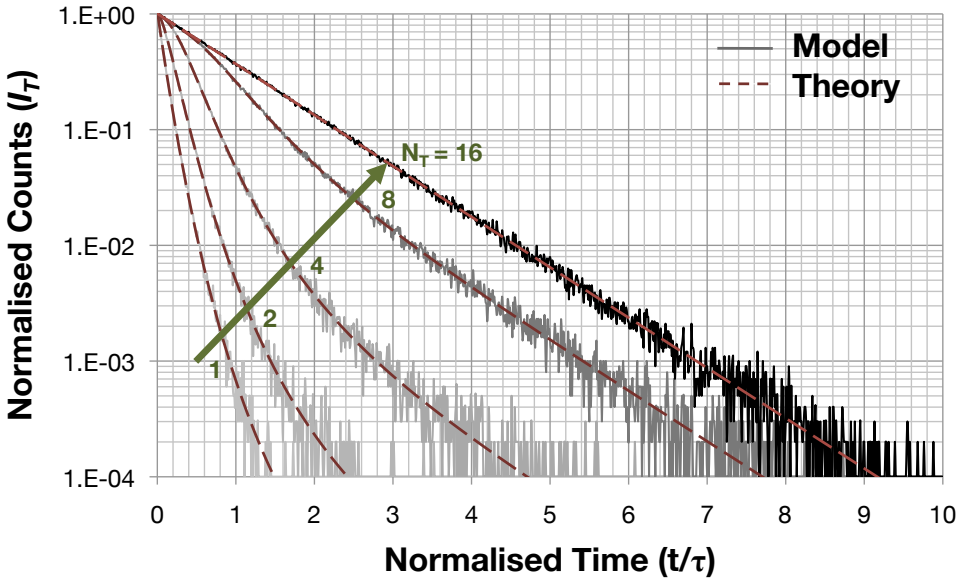


Figure 3.6: *Effect on lifetime decay of increasing the number of timing channels available per excitation period (N_T) for a fixed photon rate of $\mu = 10.0$.*

3.4.3 CMM Calculation

Due to the recursive nature of Equation 3.2, no closed form solution to the CMM calculation was found. Therefore, the graph in Figure 3.7 just shows the modelled effect of increasing μ, from 0.01 to 100.0, on the calculated lifetime (left - solid) and on the number of photons lost to pile-up (right - dashed). With the idealised detector parameters, the photons lost due to pile-up are all lost due to a lack of timing channels. The initial case of $N_T = 1$ is equivalent to the results shown in Figure 3.5. The resolvability of the CMM lifetime calculation can clearly be seen to improve for high photon-rates (μ) by increasing the number of timing channels available per excitation period, thanks to a reduction in photons lost to pile-up. A photon throughput rate of $\mu = 1.0$ appears to be achievable with four or more timing channels ($N_T \geq 4$).

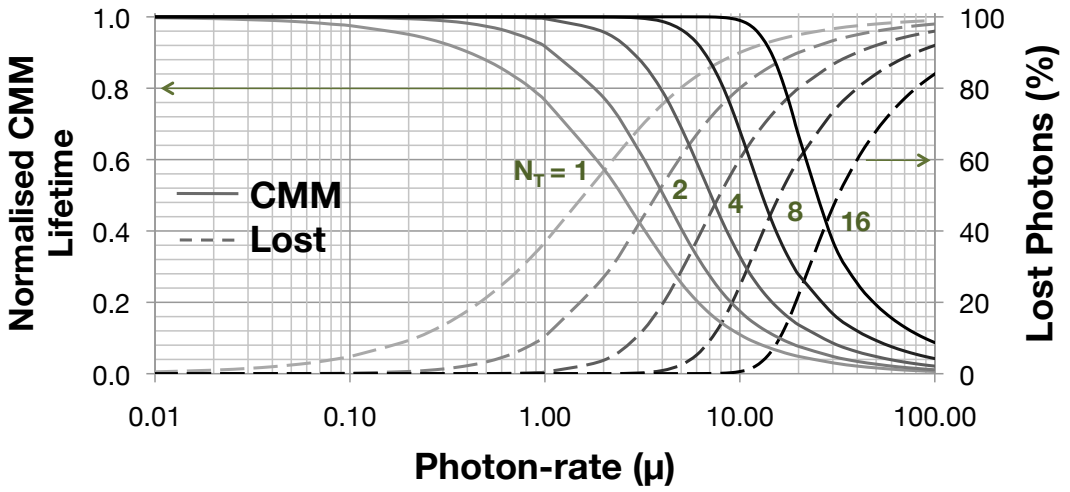


Figure 3.7: *Effect of increasing μ on lifetime calculation (solid - left) and photon loss (dashed - right) for a varying number of timing channels (N_T).*

The information in Figure 3.7 can be used to calculate the maximum available photon-rate (μ_{max}) for a given number of timing channels (N_T), assuming the fixed acceptable fluorescence lifetime calculation errors defined in Section 3.3.3. The result of this is shown in Figure 3.8 for the acceptable errors of 1 % and 5 %, where the relationship between N_T and μ_{max} is linear for $N_T \geq 4$. The results confirm that 4 timing channels are required to increase μ_{max} to over 1.0 for only 1% error in lifetime calculation using CMM, a significant improvement on classic single timing channel TCSPC set-ups. Furthermore, as also shown by Figure 3.6, 16 timing channels are required to reach $\mu = 10.0$. However, all of this analysis assumes ideal detector parameters, of which further investigation will be performed in the following sections.

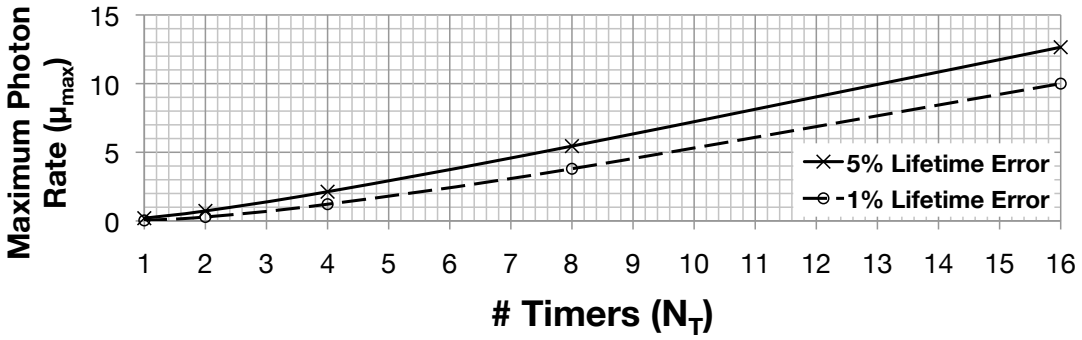


Figure 3.8: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying N_T.*

3.5 Channel Pulse-Width Pile-Up

3.5.1 Overview

In practice, the shortened pulse-width of the SiPM output will have a finite time and will therefore have a major effect on experiments. This section will investigate how this finite pulse-width affects both the captured decay histograms and the CMM fluorescence lifetime calculations. The effect that the pulse-width has on fluorescence lifetime decays is dependent not only on the pulse-width itself, but also on the lifetime being measured. Therefore, rather than describing the dependence of lifetime measurements on the absolute value of the pulse-width (t_P), it is normalised as a fraction of the lifetime (t_P/τ). Again, all other architectural variables are idealised ($N_D, N_T = \infty$ and $t_D, DCR = 0$) so that pile-up can only be caused by the parameter under investigation, the pulse-width of the SiPM (t_P).

3.5.2 TCSPC Decay Histograms

It is impossible for any event after the first to be processed within the pulse-width of the SiPM output. Therefore, the histogram for this portion of the lifetime decay ($t < t_P$) is described by Equation 2.1. The effect of the subsequent photons within a single excitation period is then to cause the decay to *recover* back towards the ideal decay histogram ($e^{-t/\tau}$). The contribution that each photon (after the first) has on the resultant histogram is given by the recursive Equation 3.3 [1] (see Appendix A.4). This equation is similar to Equation 3.2, but with the inclusion of the term e^{t_P} to describe the recovery back towards the ideal decay, after $t = t_P$.

$$I_P(\mu; t; m) = \frac{\mu}{m-1}(1 - e^{-t/\tau} \cdot e^{t_P}) \cdot I_P(\mu; t; m-1) \quad \text{for } m > 1, t \geq t_P \tag{3.3}$$

In this investigation, there is an ideal number of timing channels available ($N_T = \infty$), so all photons that do not suffer from pulse-based pile-up can theoretically be recorded. The sum of all photons within an excitation period, $\sum_{m=1}^{\infty} I_P(\mu; t; m)$, can be simplified to the bottom half of Equation 3.4 [1] (see Appendix A.4), the entirety of which describes the full decay histogram for a given photon-rate (μ) and pulse-width (t_P).

$$I_P(\mu; t) = e^{-t/\tau} \times \begin{cases} e^{-\mu\left(1-e^{-t/\tau}\right)} & t < t_P \\ e^{-\mu \cdot e^{-t/\tau}(e^{t_P/\tau}-1)} & t \geq t_P \end{cases} \tag{3.4}$$

The graph in Figure 3.9 shows the effect of increasing μ from 0.01 to 10.0 on the resulting captured TCSPC decay histogram using both the model (solid) and the theory described by Equation 3.4. In this instance the SiPM output pulse-width is set equal to the lifetime value ($t_P = \tau$), as shown by the vertical line in the figure. For time less than the pulse width ($t < t_P$), the theory and model can be seen to match exactly with the graph shown in Figure 3.4, where only a single timing channel ($N_T = 1$) is available to process photon events. For $t > t_P$, the ability to time subsequent photons can be seen by the recovery of the captured histogram towards the ideal lifetime decay. It can clearly be seen that the SiPM pulse-width has a major distorting effect on the captured histogram, at the point where $t = t_P$, and is particularly noticeable at high photon-rates.

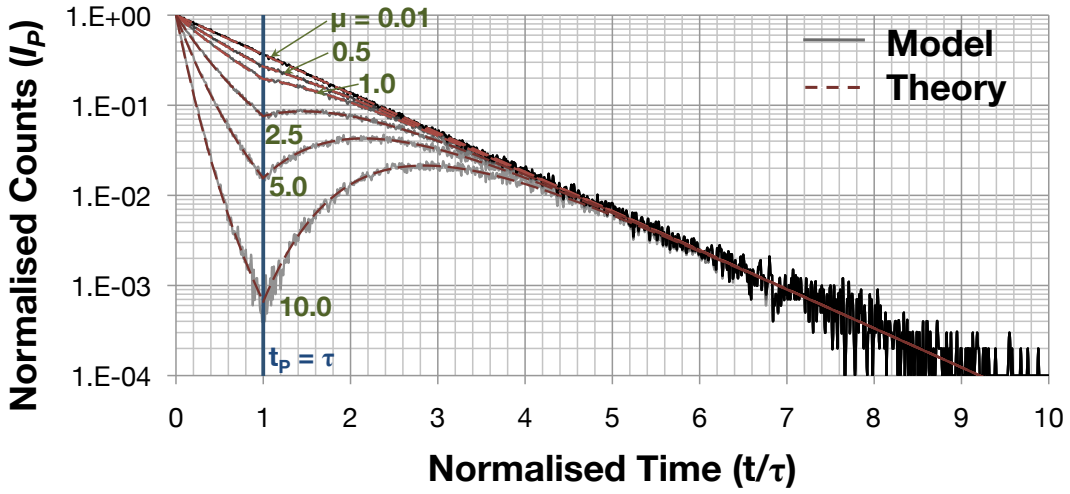


Figure 3.9: *Effect of increasing μ on the captured histogram for $t_P = \tau$ using the model (solid) and theory (dashed).*

The pulse-width to lifetime ratio (t_P/τ) can take many values depending on the process and implementation constraints (t_P), as well as the experiment being performed (τ). To investigate this, the graph in Figure 3.10 shows the effect of increasing t_P/τ from 0.02 to 4.0 on the resulting captured TCSPC decay histogram, again using both the model (solid) and the theory described by Equation 3.4 (dashed) for a fixed photon-rate of $\mu = 1.0$. The figure clearly shows that this form of pile-up is a major issue for longer t_P/τ ratios (> 0.2), whilst for shorter ratios (< 0.02) the effect is almost unnoticeable at this photon-rate. In fact, if the pulse-width is greater than the excitation period ($t_P > 1/f_E$), then this configuration will mimic a single timing channel setup with the added complexity of cyclical effects being added.

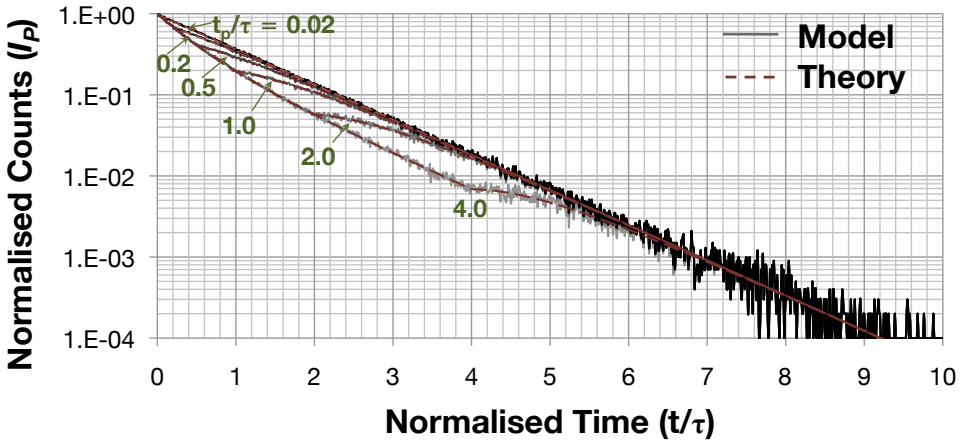


Figure 3.10: *Effect of increasing t_P/τ on the captured histogram for $\mu = 1.0$ using the model (solid) and theory (dashed).*

3.5.3 CMM Calculation

As can be expected from investigating distortion of the TCSPC histograms, t_P will have a significant effect on the fluorescence lifetime calculation by CMM. The graphs in Figures 3.11 and 3.12 show the effect of increasing μ from 0.01 to 100.0 for a range of t_P/τ ratios (from 0.005 to 5.0) for both photons lost due to SiPM pulse-width pile-up and the CMM lifetime calculation, respectively. As t_P approaches $1/f_E$ (e.g. $t_P/\tau = 5.0$), both the CMM calculation and the photons lost due to pile-up approaches the classic single timing channel TCSPC configuration, where $N_T = 1$, as shown in Figure 3.5. Furthermore, it has the added disadvantage of introducing cyclic effects as the pulse-width will cause events at the beginning of the next excitation period to be missed.

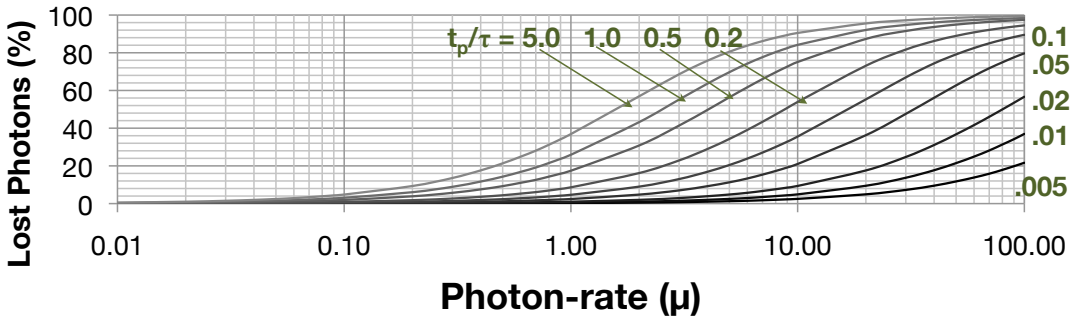


Figure 3.11: *Effect of increasing μ on photons lost due to SiPM pulse-width pile-up for varying pulse-width to lifetime ratios (t_P/τ).*

Reducing t_P/τ below this worst case scenario causes the CMM calculation to improve towards a factor of 1.0, before continuing to over-estimate up to a ratio of $t_P/\tau \approx 0.5$, as shown in the graph at the top of Figure 3.12. The CMM calculation appears to extend the range of resolvability beyond $\mu = 10.0$ for a t_P/τ ratio of ≈ 1.5, however this is a false-positive result as many photons are lost in this instance and the lifetime is only calculated correctly as the pulse-width does not affect the centre of mass of the histogram. Continuing to reduce t_P/τ below 0.5 then causes the CMM calculation to approach the ideal factor of 1.0 due to very few photons being lost to pile-up, as shown at the bottom of Figure 3.12.

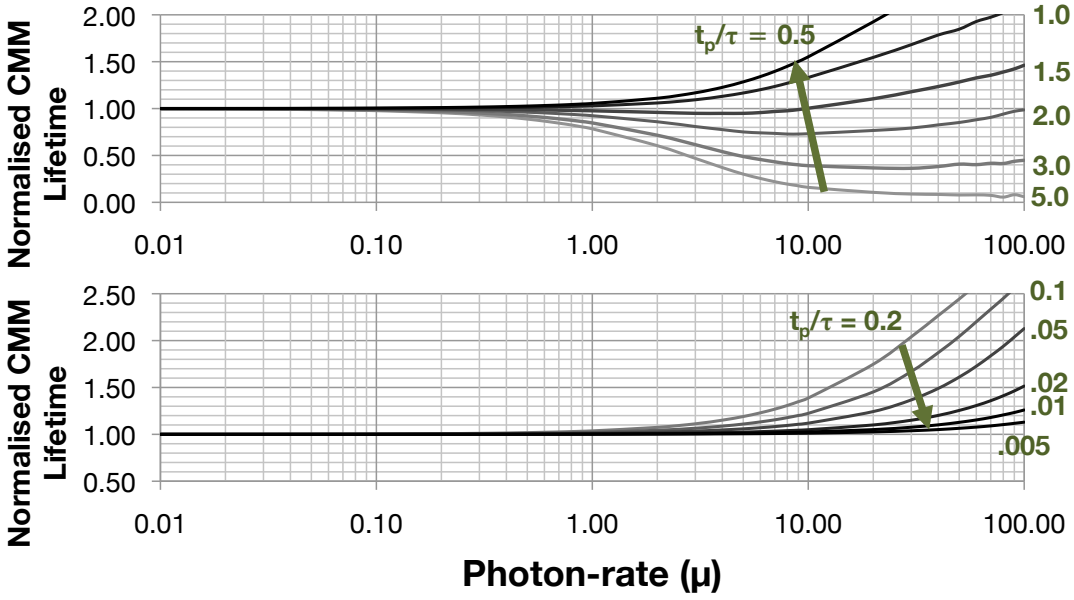


Figure 3.12: *Effect of increasing μ on lifetime calculation for varying pulse-width to lifetime ratios (t_P/τ).*

Similarly to Figure 3.8, the maximum available photon-rate (μ_{max}) for a given SiPM pulse-width to lifetime ratio (t_P/τ) and acceptable fluorescence lifetime calculation errors of 1 % and 5 % is given in Figure 3.13. The split in the graph is caused by the case mentioned above, where at around $t_P/\tau = 1.5$, the resolvability returns a false-positive result at high μ. The results show that t_P/τ should be below 0.1 so that this situation does not occur. A value of $t_P/\tau = 0.05$ provides a photon throughput of $\mu_{max} > 1.0$ for only 1 % error in lifetime calculation. It is clear from the simulation results that the SiPM pulse-width should be as short as possible, given the implementation and process constraints, to allow the maximum photon-rate (μ_{max}) to be as high as possible.

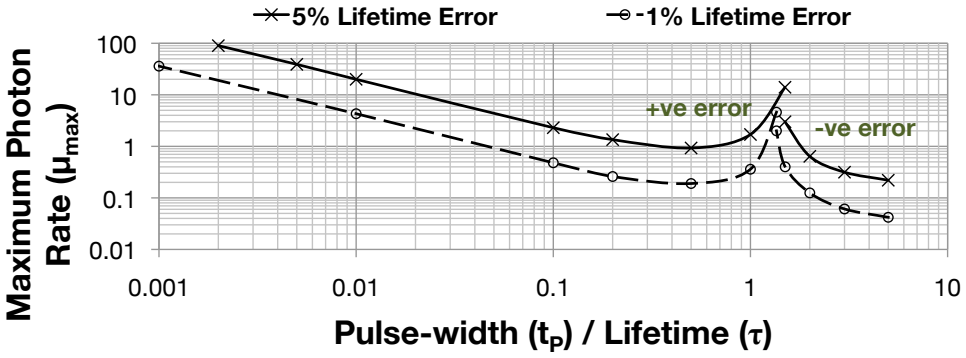


Figure 3.13: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying t_P/τ.*

3.6 Detector Pile-Up

3.6.1 Overview

The final architectural parameters to investigate are those concerning the detectors within the SiPM. This section will look at the combination of the number of detectors (N_D), their dead-time (t_D) and their dark count rate (*DCR*). For the same reasons as in Section 3.5, the dead-time is given as a fraction of the lifetime (t_D/τ). As explained in Section 3.2.4, the *DCR* will be assumed to be a fixed value for each detector in the configuration (DCR_D), so the total *DCR* is equal to $N_D \cdot DCR_D$. Furthermore, the other architectural parameters are assumed to be ideal ($N_T = \infty$ and $t_P = 0$), so that all pile-up is caused by the detector parameters. Due to the increased number of varying parameters, theory is not used in this investigation.

3.6.2 TCSPC Decay Histograms

The graphs in Figure 3.14 show the effect of increasing the number of detectors (N_D) on the resulting TCSPC decay histogram for both a long $t_D/\tau = 8.0$ (top) and a short $t_D/\tau = 1.0$ (bottom), shown by the vertical lines, and with a fixed DCR_D of 1 kHz and photon-rate (μ) of 5.0. As expected, increasing the number of detectors (N_D) allows the histogram to approach the ideal decay, in a similar way to increasing the number of timing channels (N_T), as was shown in Figure 3.6. However, the rate at which this improvement back to the ideal decay is made is much slower for increasing N_D than for increasing N_T, requiring $N_D > 32$, for a smaller μ, before the decay is approximately ideal.

The graph at the top of Figure 3.14 shows the resulting histograms for varying N_D with a fixed $t_D/\tau = 8.0$ and μ = 5.0. The long dead-time produces very prominent cyclic effects, as highlighted by the inset in the graph. For the single detector configuration ($N_D = 1$), the peak counts in the histogram (C_P) have been reduced by over 50 %. Furthermore, the effect of t_D/τ is very similar to the effect of t_P/τ, causing a significant loss of photons around the dead-time. This is most clear from the bottom graph in Figure 3.14, where the distortions caused by the single detector configuration ($N_D = 1$) resembles the shape of the decays shown in Figures 3.9 and 3.10.

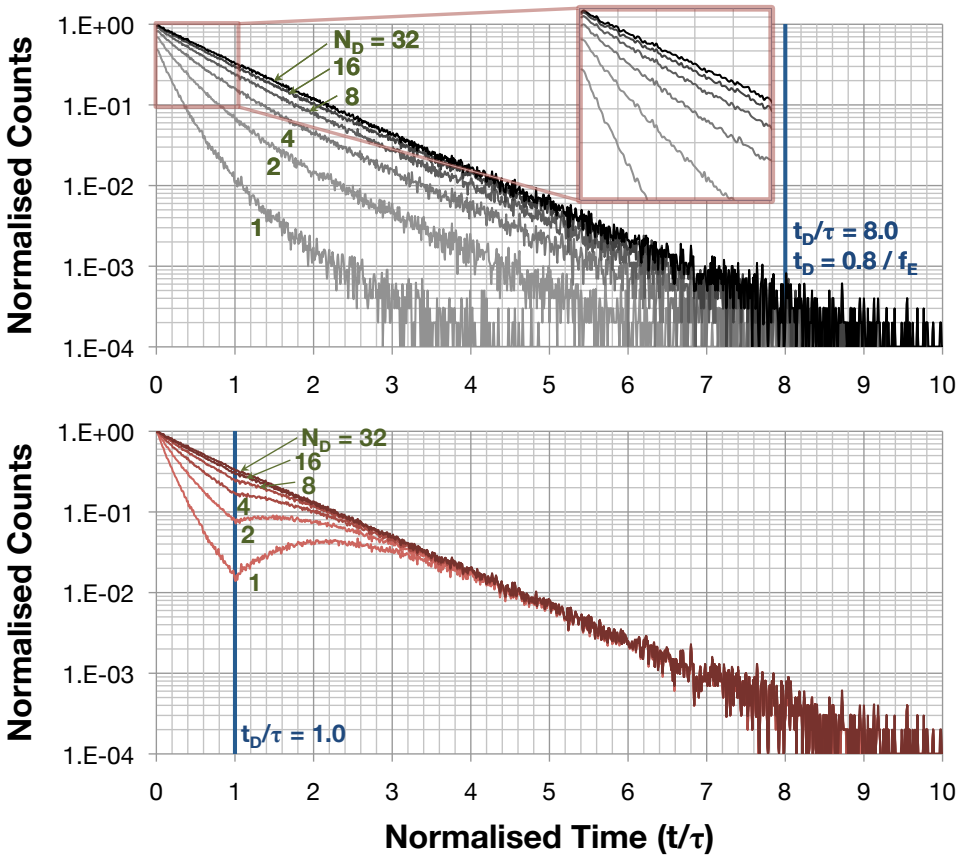
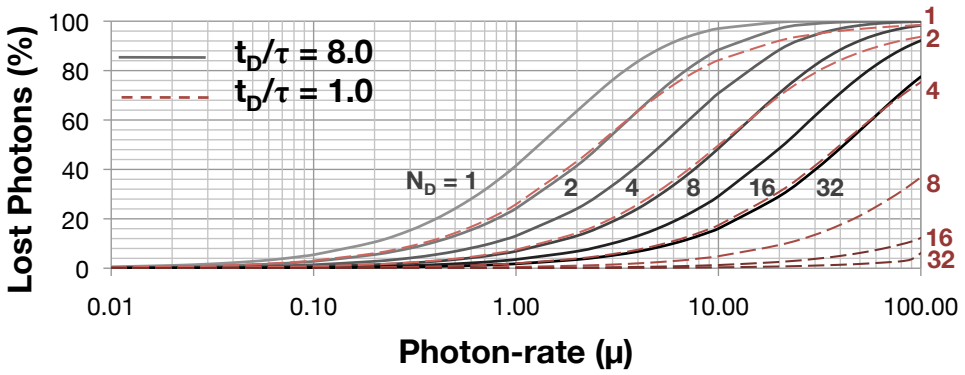


Figure 3.14: *Effect of increasing N_D on the captured histogram for $t_D/\tau = 8.0$ (top) and $t_D/\tau = 1.0$ (bottom), $\mu = 5.0$ and DCR_D of 1 kHz.*

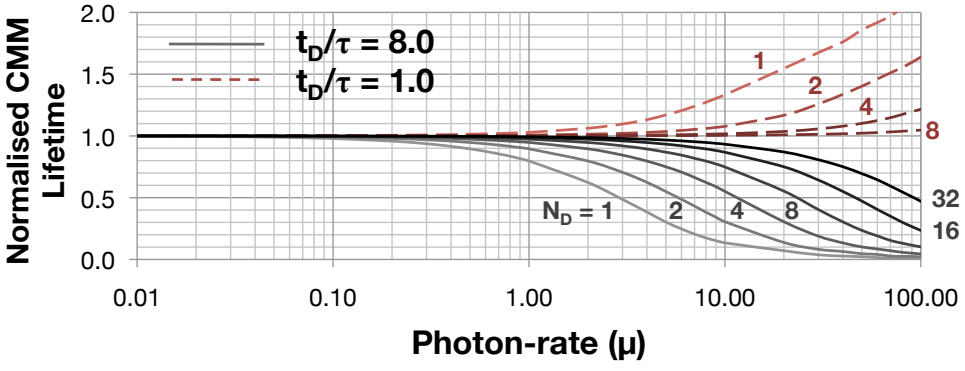
Finally, even though a representatively large DCR_D value of 1 kHz is chosen for the results shown in Figure 3.14, the contribution this makes to the histograms is negligible and cannot be seen. This is explained by the high signal to noise ratio (SNR) in this configuration, caused by such a high photon-rate (μ) of 5.0, which is a photon throughput of 100 MHz for a typical excitation rate of 20 MHz.

3.6.3 CMM Calculation

The graphs in Figures 3.15a and 3.15b show the effect of increasing the photon-rate (μ) using varying numbers of detectors (N_D) on the photons lost due to detector pile-up and on the fluorescence lifetime calculation using CMM, respectively. In each case, dead-time to lifetime ratios (t_D/τ) of 8.0 and 1.0 are simulated and shown using solid and dashed curves, respectively. The similarities with SiPM pulse-width are apparent (see Figures 3.11 and 3.12), where longer dead-times cause comparatively more losses and give negative lifetime calculation errors whilst shorter dead-times cause comparatively less losses and give positive lifetime calculation errors. It is clear to see from these results that for longer t_D, more detectors are required to reach an equivalent lifetime calculation accuracy than for short t_D. It is expected that t_D/τ ratios will be $>$ 1.0, given typical passively quenched SPAD dead-times of tens of nanoseconds and common organic fluorescence lifetime probe values below 20 ns.



(a)



(b)

Figure 3.15: *Effect of increasing μ on* **(a)** *photons lost due to detector pile-up and* **(b)** *the lifetime calculation, for varying number of detection elements (N_D) and dead-times (t_D/τ) of 8.0 (solid) and 1.0 (dashed).*

The maximum available photon-rate (μ_{max}) for a given number of detectors (N_D) and acceptable fluorescence lifetime calculation errors of 1 % and 5 % is shown in Figure 3.16. The results are given for t_D/τ = 1.0 and t_D/τ = 8.0, with the shorter dead-time requiring just 3 detectors to achieve a maximum photon-rate (μ_{max}) of over 1.0 for a 1 % error, whilst the longer dead-time requires $\approx$ 20 detectors for the same constraints. Relaxing the acceptable error to 5 % allows photon-rates (μ_{max}) of up to 5.0 for the same N_D and t_D/τ configurations.

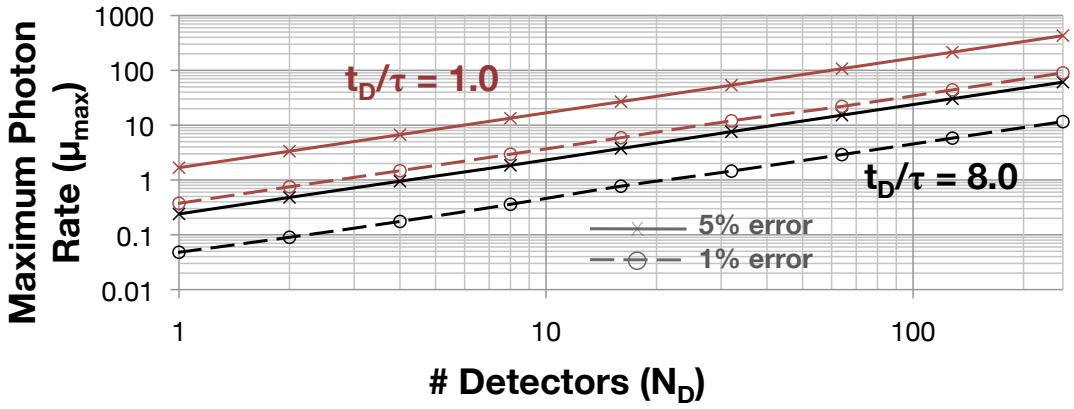


Figure 3.16: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying N_D.*

3.7 Combined Effects

3.7.1 Overview

The three main architectural parameters have been investigated individually in Sections 3.4, 3.5 and 3.6, in each case idealising the parameters not under study. This section will now look at the combined effects of all parameters simultaneously. These parameters are the number of detectors (N_D), detector dead-time (t_D), dark count rate (*DCR*), channel pulse-width (t_P) and the number of timers per excitation (N_T).

As concluded in sections 3.5 and 3.6, the channel pulse-width and detector dead-time will be designed to be as short as possible given the process and implementation constraints. Therefore, to reduce the parameter set for this investigation, these values will be fixed based on expected values from previous similar work. The values chosen for this investigation are t_D = 20 ns and t_P = 500 ps, based on the values of a similar approach in a slower process [114]. Furthermore, the lifetimes under observation will be set to τ = 5 ns and 20 ns to provide representative results

for typical organic and inorganic fluorophore lifetimes. These real values translate to modelled parameters of t_P/τ = 0.025 and 0.1, with the detector dead-time now being represented by $t_D = 40 \cdot t_P$. These values are shown to be short enough to provide photon-rates in excess of the excitation rate ($\mu > 1.0$), as can be seen in Figures 3.13 and 3.16.

This leaves the number of timers and the number of detectors (with their added *DCR*) to be varied in this investigation. As with section 3.6, it will be assumed that the *DCR* is a fixed equal value of DCR_D = 1 kHz for each detector. Although an over-estimate, this value will be insignificant for the photon throughputs to be investigated – which are in the region of tens to hundreds of MHz (assuming an excitation frequency, f_E, of tens of MHz) – even for large numbers of detectors. This section will continue by looking at a typical histogram given the parameters presented above, before looking at varying the number of timers (N_T) and then the number of detectors (N_D) individually.

3.7.2 TCSPC Decay Histograms

The graph in Figure 3.17 shows a representative TCSPC histogram (black) for $t_P/\tau = 0.1$, $N_T = 4$, $N_D = 16$ and $\mu = 1.0$. The values for the number of timers and detectors are chosen based on the required number to provide a photon-rate in excess of the excitation frequency ($\mu > 1.0$) for a lifetime calculation error of under 1 %, as shown in Figures 3.8 and 3.16. Furthermore, the longer lifetime case, where $t_P/\tau = 0.025$ is not shown due to the minimal effect such a short pulse-width has on the histogram at this photon rate. The effect that the pulse-with (t_P) makes on the histogram is seen at $t/\tau = 0.1$ and is highlighted by the inset in the graph, whilst the contribution that each timer (n_T) makes to the total histogram is also shown (grey).

3.7.3 Timing Channel Pile-Up

Similarly to Section 3.4.3, CMM lifetime calculations are performed for varying both the photon-rate (μ) between 0.01 and 100.0 and the number of timers per excitation (N_T) between 1 and 16. However, in this case the remaining parameters are set to the non-ideal values introduced in section 3.7.1. Furthermore, the number of detectors (N_D) is fixed at 16, based on the required number to provide a photon-rate in excess of the excitation frequency ($\mu > 1.0$) for a lifetime calculation error of under 1 %, as shown in Figure 3.16.

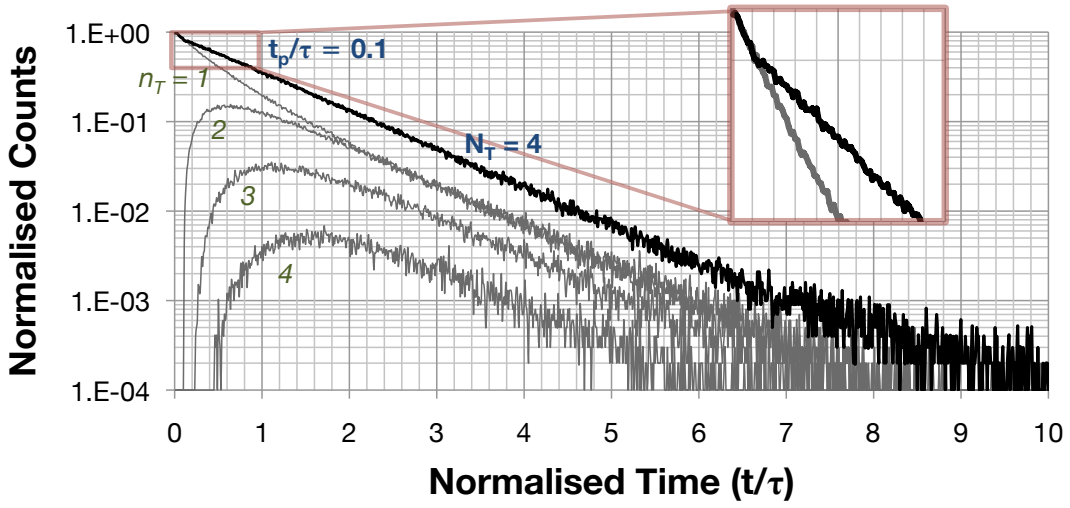
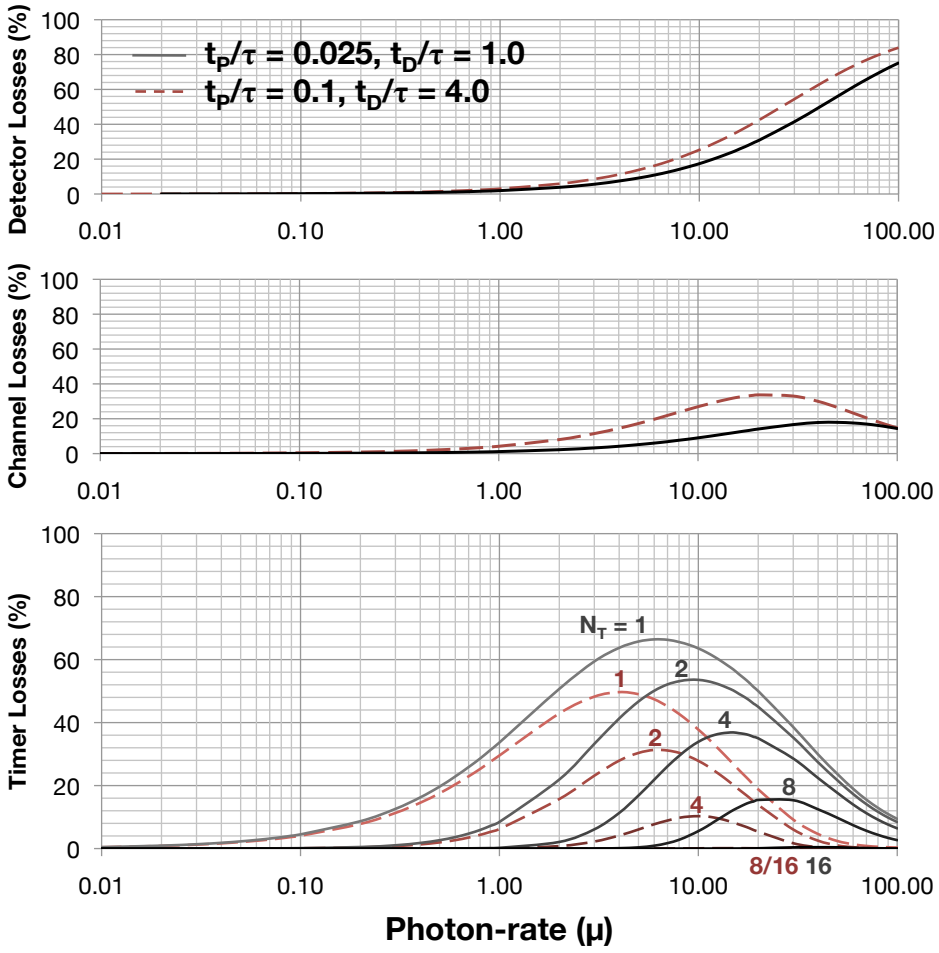


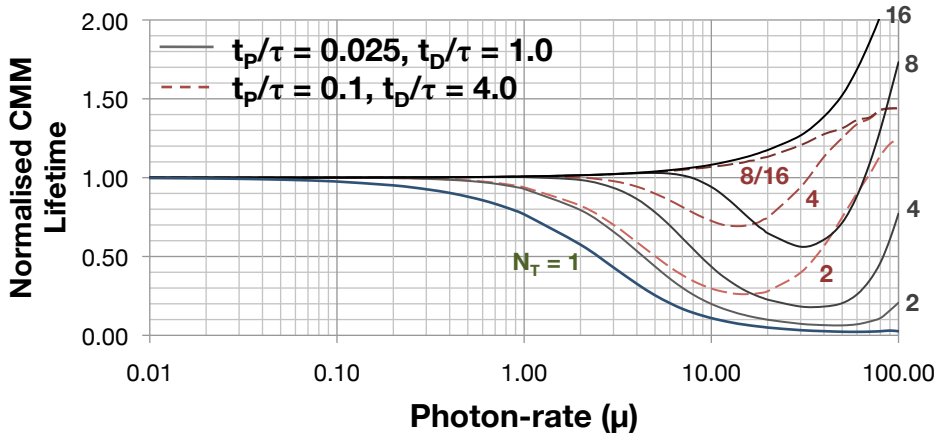
Figure 3.17: *Effect of combined (non-ideal) parameters ($t_P/\tau = 0.1$, $N_T = 4$ and $N_D = 16$) on the lifetime decay (black) for a fixed photon rate of $\mu = 1.0$, showing contribution of each timer, n_T (grey).*

The graphs in Figure 3.18a show the percentage of photons lost due to each form of pile-up (detector, channel and timer) as a function of the photon-rate and for different numbers of timers. As expected, the photons lost due to detector and channel pile-up are completely independent of the number of timers, as shown by the top two graphs in the figure. Being the first point of pile-up, the detector losses match well with the results shown in Figure 3.15a. However, the channel losses are significantly reduced from Figure 3.11 due to the losses that have already occurred at the detectors. Moreover, the combined effect of detector and channel losses mean that less photons are available to be processed, in turn decreasing the timer losses at high photon-rates ($\mu > 10.0$) in comparison to the previous results in Figure 3.7. In fact, by $\mu = 100.0$, timing losses are almost zero, even for a single timer ($N_T = 1$). The effect that the lifetime (τ) has on the photons lost is shown by the dashed ($\tau = 5$ ns) and solid ($\tau = 20$ ns) lines in the graphs. Due to the additional photons lost at the detector and channel for the shorter lifetime, less photons actually make it to the timers to be processed and so here the photon losses are actually reduced for the shorter lifetime.

The CMM lifetime calculation results are shown in the graph in Figure 3.18b. Both the longer (solid) and shorter (dashed) lifetime results show a clear transition between the CMM calculation being limited by a lack of timers ($N_T = 1$) and being limited by the detectors and channel pulse-width ($N_T = 16$), where the former produces negative errors and the latter produces positive errors. This result is backed up by the timer photon losses shown in the



(a)



(b)

Figure 3.18: *Effect of increasing μ on* **(a)** *each form of photon loss and* **(b)** *the lifetime calculation, for a varying number of timers (N_T) and pulse-widths (t_P/τ) of 0.1 (solid) and 0.025 (dashed).*

graph at the bottom of Figure 3.18a, where there are zero timer losses for $N_T = 16$ (as well as $N_T = 8$ for the shorter lifetime), independent of the photon-rate. Furthermore, the longer lifetime results show a much higher positive error in CMM calculation for $\mu > 10.0$ when there are sufficient timers to process *all* of the late photon arrivals of a severely distorted histogram that will be similar to the one shown in Figure 3.9.

Finally, the maximum available photon-rate (μ_{max}) for a given number of timers (N_T) is given in Figure 3.19, where the top graph shows the results for the longer lifetime ($t_P/\tau = 0.025$) and the bottom graph for the shorter ($t_P/\tau = 0.1$). Due to the combined effects of the detector dead-time and channel pulse-width, CMM lifetime estimates can have either a positive or negative error depending on the photon-rate and the lifetime under observation. These positive and negative errors can be seen graphically in Figure 3.18b, producing a *split* in the graphs in Figure 3.19. Even with the combined non-ideal parameters, $N_T = 4$ still provides a throughput of over 1.0 for both lifetimes simulated. Once detector and channel losses begin to dominate and the error goes positive, the benefit of adding timers is reduced and the maximum resolvability flattens out at $\mu_{max} \approx 1.5$. However, for longer lifetimes, by increasing the number of timers, much higher throughputs are possible (e.g. for $N_T = 8$, $\mu_{max} = 10.0$) at the expense of an increased error in lifetime measurement of 5 %.

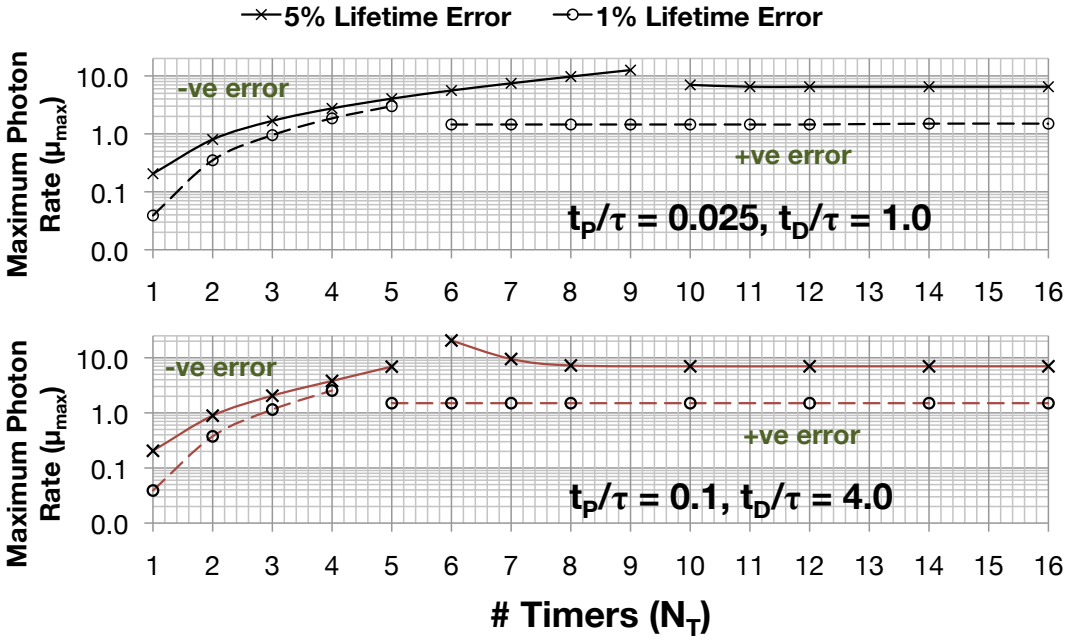


Figure 3.19: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying N_T.*

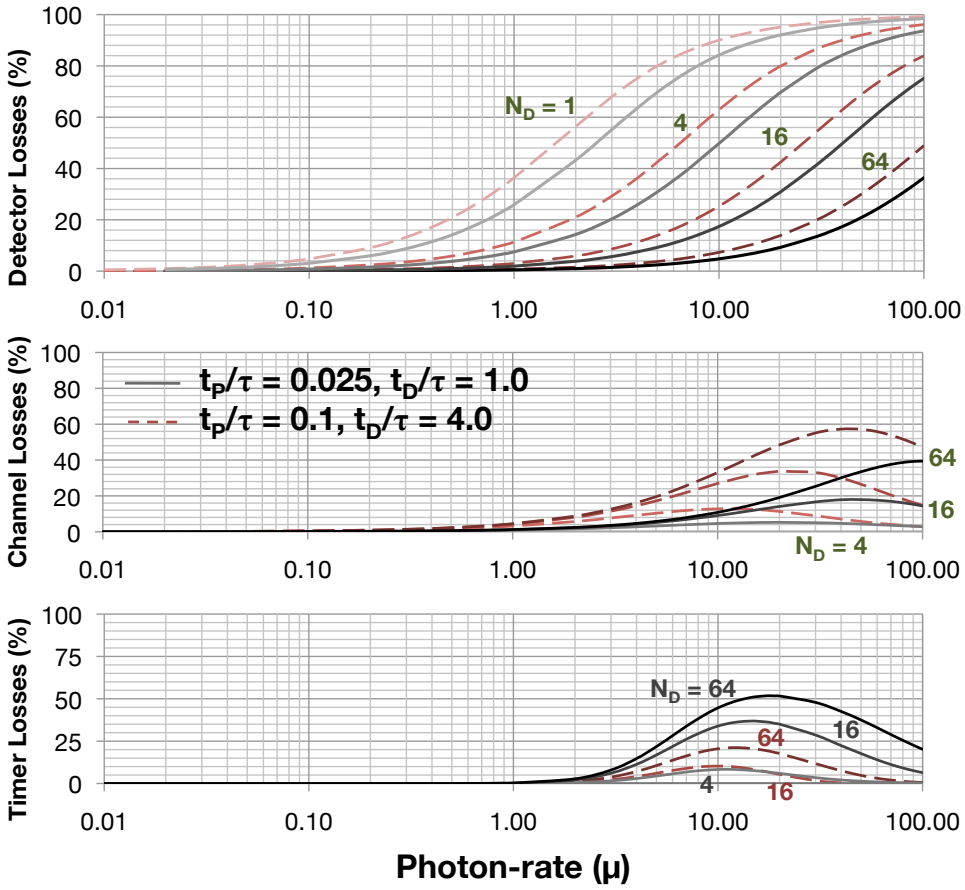
3.7.4 Detector Pile-Up

Next, CMM lifetime calculations are performed for varying both the photon-rate (μ) between 0.01 and 100.0 and the number of detectors between 1 and 1024, whilst the remaining parameters are fixed as introduced in section 3.7.1. The number of timers (N_T) is fixed at 4, based on the required number to provide a photon-rate in excess of the excitation frequency ($\mu > 1.0$) for a lifetime calculation error of under 1 %, as shown in Figure 3.8.

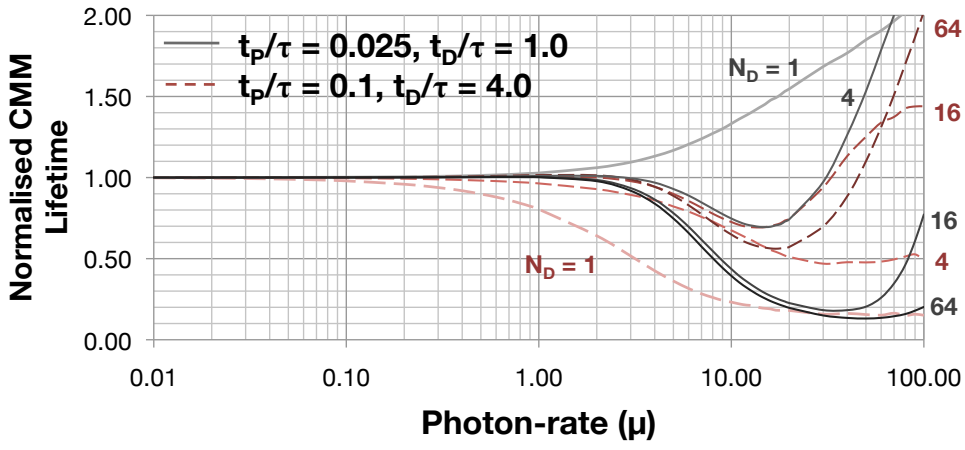
The graphs in Figure 3.20a show the percentage of photons lost due to each form of pile-up (detector, channel and timer) as a function of the photon-rate and for different numbers of detectors. As expected, the detector losses closely match the results from Figure 3.15a. For a low number of detectors ($N_D \leq 4$), photon losses are almost completely dominated by detector pile-up, with channel and timer losses being limited to below 10 % each for both lifetimes. As N_D is increased, detector losses reduce but are replaced with a mix of channel and timer losses. For the parameters chosen, these channel and timer losses dominate in different regions of photon-rate, with channel losses being relatively more spread-out than timer losses, which are more concentrated around $\mu = 10.0$.

The CMM lifetime calculation results are shown in the graph in Figure 3.20b. These results follow the same pattern as first presented in Figure 3.15b, where shorter lifetimes (long t_P and t_D) give negative errors and longer lifetimes (short t_P and t_D) give positive errors, with both tending towards a correct calculation as N_D is increased. The added complexity in this case is the inclusion of a finite number of timers ($N_T = 4$), which further pulls the lifetime calculation negative. This is more noticeable for the longer lifetime results, due to losses being dominated by timer pile-up, whereas the shorter lifetime is dominated by channel pile-up, as shown in Figure 3.20a.

Finally, the maximum available photon-rate (μ_{max}) for a given number of detectors (N_D) is given in Figure 3.21, where the top graph shows the results for the longer lifetime ($t_P/\tau = 0.025$) and the bottom graph for the shorter ($t_P/\tau = 0.1$). As with Figure 3.19, due to the combined effects of the detector dead-time and channel pulse-width, CMM lifetime estimates can have either a positive or negative error depending on the photon-rate and the lifetime under observation. In this case however, for a 5 % CMM lifetime calculation error, the shorter lifetime provides a negative error independent of N_D. This can be seen in both Figure 3.20b and by the lack of a disjoint in the solid curve (5 %) of the lower graph in Figure 3.21.



(a)



(b)

Figure 3.20: *Effect of increasing μ on* **(a)** *each form of photon loss and* **(b)** *the lifetime calculation, for a varying number of detectors (N_D) and pulse-widths (t_P/τ) of 0.1 (solid) and 0.025 (dashed).*

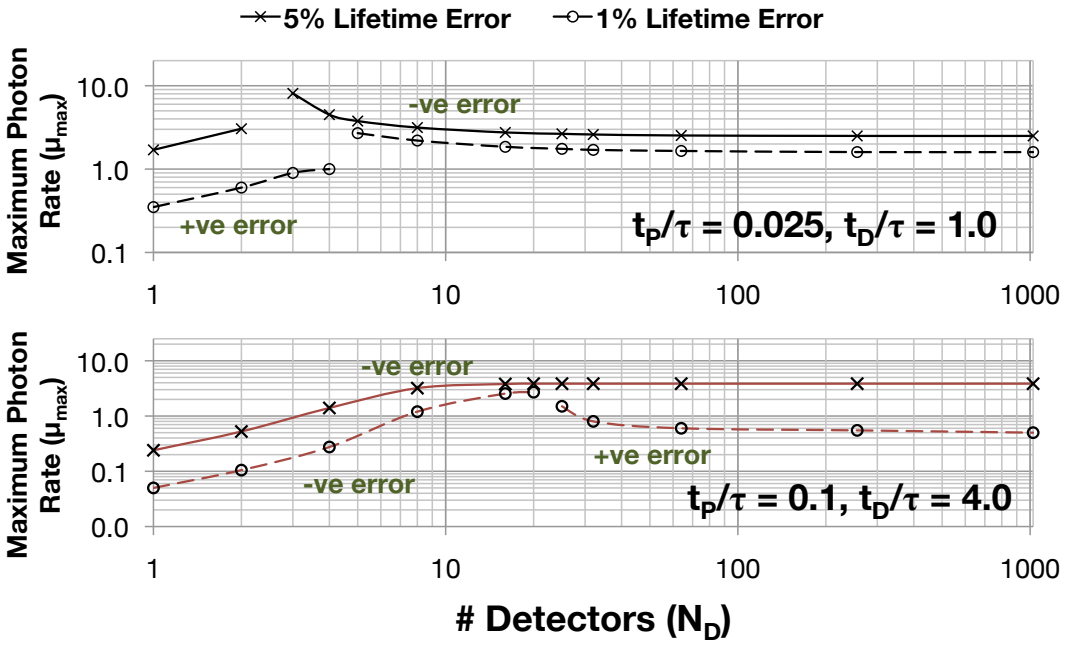


Figure 3.21: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying N_D.*

From these graphs, it is clear to see that the ideal number of detectors lies between $N_D \approx 8$ and $N_D \approx 20$, where a photon-rate (μ) of over 1.0 is possible for for a CMM lifetime calculation error of less than 1 %. Furthermore, it can be seen that increasing N_D beyond 20 provides no additional photon-rate gains due to the pile-up losses being limited by the number of timers. In fact, for $N_D > 30$ the shorter lifetime results reduce the available throughput below $\mu = 1.0$ for a 1 % error in lifetime calculation. It is clear therefore that simply increasing N_D does not necessarily produce positive results.

Further investigation of the relationship between N_D and N_T is necessary to make a sensible decision on the parameters to select for implementation. This will be introduced in the following section along with a more thorough investigation of the effect of the lifetime (τ), detector dead-time (t_D) and channel pulse-width (t_P) have on these chosen parameters.

3.8 Architecture Proposal

3.8.1 Overview

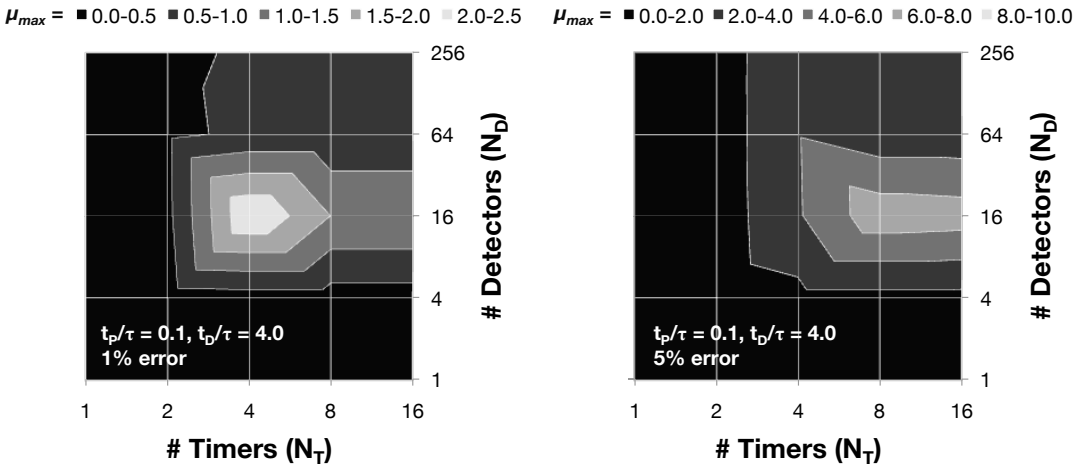
Using the theoretical and simulated results presented so far in Chapter 3, this section will introduce proposals for the architectural parameters to enable the design of a fluorescence lifetime sensor capable of achieving a photon throughput in excess of the excitation frequency ($\mu > 1.0$). A final investigation of the relationship between the controllable parameters of the device architecture – the number of timing channels (N_T) and the number of detectors (N_D) – is presented to make suitable choices for their values. The channel pulse-width (t_P), detector dead-time (t_D) and detector *DCR* will take the values introduced in Section 3.7.1. Once the individual architecture proposals have been presented, they will be used to model the system with a varying lifetime (τ) to highlight the expected device performance and its limitations.

3.8.2 Parameter Selection

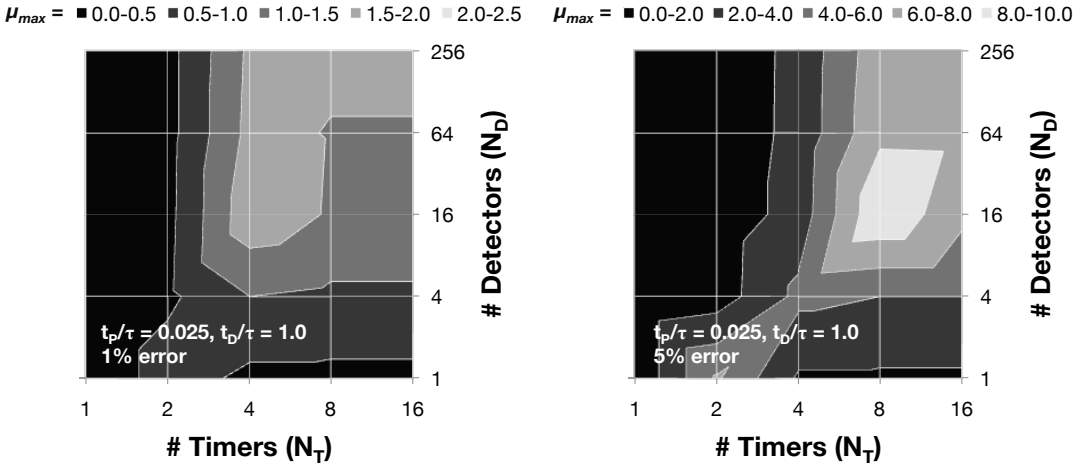
The parameters that can be varied in the sensor design – the number of detectors (N_D) and the number of timers (N_T) – have been investigated individually to determine their response to increasing the photon throughput (μ) and varying the lifetime (τ) being measured. However, the results in Figures 3.19 and 3.21 are insufficient to make an informed decision on these parameters for the final design. Therefore to obtain a better understanding of their relationship, a final investigation is performed by sweeping both N_D and N_T together for the same t_P/τ values introduced in Section 3.7.1. The results from this investigation for both short and long lifetimes are shown in the two dimensional surface plots in Figures 3.22a and 3.22b, respectively, where the maximum available photon throughput is shown for 1 % (left) and 5 % (right) error in lifetime calculation.

The short lifetime with 1 % error results in a clear peak μ_{max} at $N_T = 4$ and $N_D = 16$, which are the parameters that have been presented throughout Section 3.7. However, this available throughput of over $\mu = 2.0$ is only achievable as the detector and timer losses balance each other out at these values, as is shown in Figures 3.18 and 3.20, which extends the apparent resolvability. The 5 % error chart therefore provides a better understanding of the relationship between the parameters, where the number of detectors is also optimal at $N_D = 16$. Increasing the number of timing channels however appears to extend the maximum throughput to $\mu_{max} > 6.0$ with $N_T = 8$.

Due to the shorter relative channel pulse-width (t_P), the longer lifetime results show an increasing maximum available throughput by increasing the number of detectors, for both 1 % and 5 % errors in calculation. Furthermore for a 5 % error, setting $N_T = 8$ and $N_D = 16$ provides a maximum available photon throughput of $\mu_{max} = 9.8$, which is almost 50 times higher than classical pile-up limited single channel TCSPC, which can only provide $\mu_{max} = 0.2$ for the same error, as shown in Section 3.3.3.



(a)



(b)

Figure 3.22: *Maximum available photon-rate (μ_{max}) for* **(a)** *short lifetime and* **(b)** *long lifetime, and for 1 % (left) and 5 % (right) lifetime calculation error by varying both N_T and N_D.*

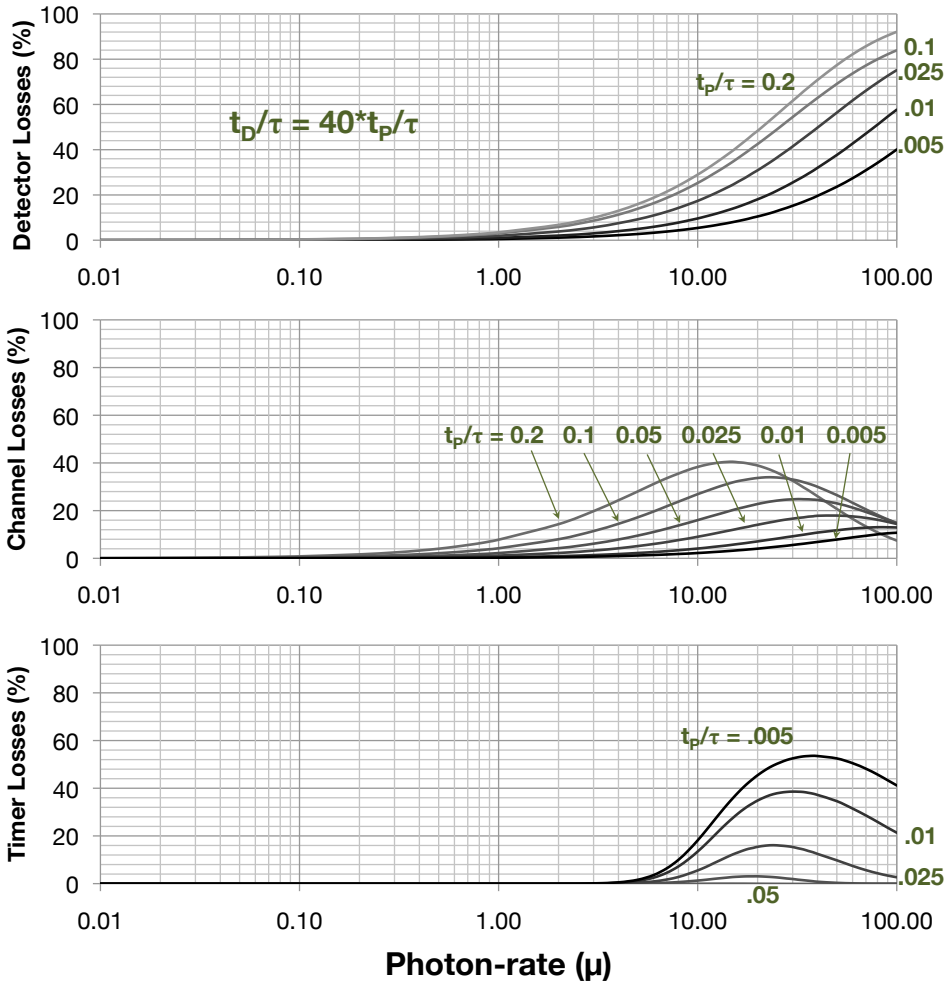
The relationship between the number of timers (N_T) and the number of detectors (N_D) is of particular interest to direct the choice of parameters for implementation of the chosen architecture. It may be initially surprising that many more detectors than timers are required to achieve a comparable throughput increase ($N_D \gg N_T$), despite both the individual detection and timing elements only being capable of processing at most one photon per excitation cycle (assuming $t_D \approx \tau$). However, photon events are distributed randomly amongst the N_D detectors, whereas the distribution of events to the N_T timers is deterministic. Therefore, the relative variance in the number of photons per timer is smaller than that for the detectors ($\sqrt{N_T}$ and N_D, respectively) [1].

Using the relationship described above, it would seem sensible to select values of $N_D = 16$ and $N_T = \sqrt{N_D} = 4$. However, as the results in this chapter so far (including the graphs in Figure 3.22) show, the inclusion of a non-ideal channel pulse-width (t_P) suggests that optimal performance is achieved by increasing the number of timers (N_T) slightly beyond $\sqrt{N_D}$. This is particularly noticeable when longer lifetime fluorophores are modelled, where this architecture appears to be best suited. Therefore based on this information, parameters of $N_D = 16$ and $N_T = 8$ are chosen for the final architecture.

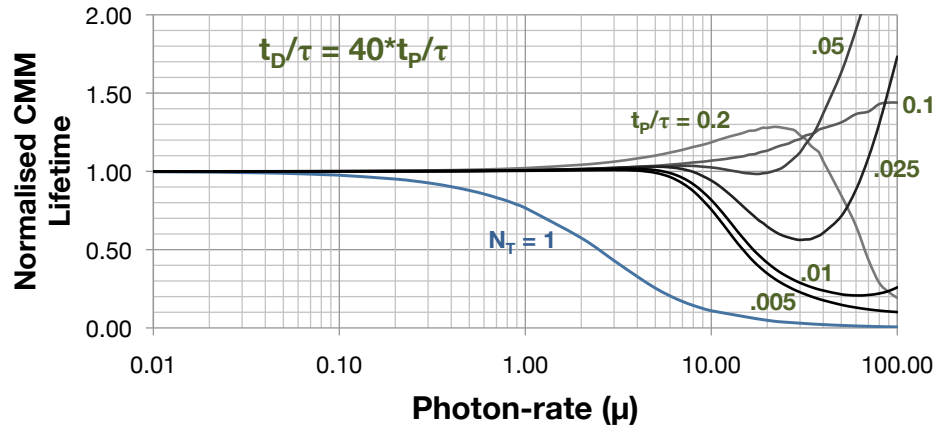
3.8.3 Simulating Proposed Parameters

The model is now used to simulate the expected performance of the chosen controllable parameters of $N_T = 8$ and $N_D = 16$, for varying lifetime values (τ). As introduced in Section 3.7.1, the lifetime is varied by providing a different t_P/τ ratio. Furthermore, t_D is assumed to be $40 \cdot t_P$. The values of t_P/τ are varied from 0.2 to 0.005, which represent real lifetimes of $\tau = 2.5$ ns to 100 ns, respectively, assuming $t_P = 500$ ps. A pessimistic *DCR* of 1 kHz per detector is chosen, which provides 16 kHz from the entire SiPM.

The graphs in Figure 3.23a show the percentage of photons lost due to each form of pile-up (detector, channel and timer) as a function of the photon-rate and for different t_P/τ ratios. The channel losses clearly increase the fastest with photon rate, beginning to be noticeable as low as $\mu = 1.0$ and growing to 40 % by $\mu = 10.0$ for the shortest lifetime. As detector losses begin to dominate at higher photon rates, a smaller portion of photons are output from the SiPM, so both channel and timer losses appear to fall. For $t_P/\tau > 0.05$ ($\tau < 10$ ns), timer losses are effectively zero as the losses that have occurred at the detector or in the channel are so high that the timing channels are never saturated.



(a)



(b)

Figure 3.23: *Effect of increasing μ on* **(a)** *each form of photon loss and* **(b)** *the lifetime calculation, for varying t_P/τ.*

The CMM calculation results are shown in Figure 3.23b, where the single channel case is shown as a comparison to highlight the improvements available with this parameter selection. As expected from the photon losses, longer lifetimes are dominated by timer loss, so provide a negative error. Conversely shorter lifetimes are dominated by detector and channel losses, so provide a positive error. This positive error is particularly noticeable at $t_P/\tau = 0.2$ ($\tau =$ 2.5 ns), where the error appears quite significant at $\mu > 1.0$. Decreasing t_P/τ below 0.2 appears to extend the lifetime resolvability up to photon rates of $\mu \approx 5$ in all cases, for an improved error. This highlights the capability of the device to perform better with longer lifetimes.

Finally, the maximum available photon rate (μ_{max} for a given t_P/τ ratio is shown in Figure 3.24. Again, these results are shown for a 1% (dashed) and 5% (solid) error in lifetime calculation and the expected performance of a single channel TCSPC arrangement is shown for comparison (blue). For $t_P/\tau < 0.1$ ($\tau > 5$ ns), the architecture is capable of achieving a photon throughput in excess of the excitation frequency (shown by the solid red line in the figure), and $\mu_{max} > 10.0$ is possible in some cases for 5 % calculation error. The limitation of the architecture is apparent when $t_P/\tau > 0.1$ ($\tau < 5$ ns), where the maximum throughput drops below the excitation frequency for a 1 % error in lifetime calculation. However, the performance in relation to the lifetime in real terms is very dependent on the channel pulse-width (t_P), so reducing this as low as possible is a key factor in the design of the SiPM, which will be presented in Section 4.3.

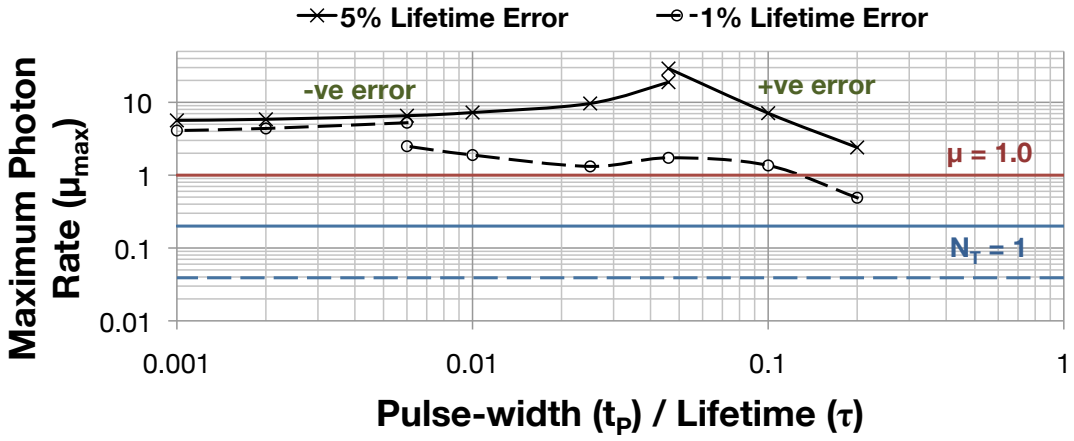


Figure 3.24: *Maximum available photon-rate (μ_{max}) for 1 % (dashed) and 5 % (solid) lifetime calculation error by varying t_P/τ.*

3.9 System Precision

Up to now, all simulations have been performed using $C_P = 10,000$ to minimise statistical error whilst studying the effects of photon throughput (in terms of μ) on the *accuracy* of lifetime calculation using CMM with different architecture configurations. This section will now investigate varying C_P – and hence the *total* number of photon counts (N_C) – to quantify the *precision* of the CMM calculation within the proposed integrated system architecture ($N_T = 8$ and $N_D = 16$). This investigation will assume a fixed excitation frequency (10 MHz) and experimental acquisition time (1 ms), so μ will be varied proportionately to N_C.

The graph in Figure 3.25 shows the results of this investigation for different DCR levels of 0 Hz (ideal shot noise limited) and 16kHz (16×1 kHz). With 1 ms exposures, this results in a noise level of ≈ 16 counts. As expected, the results are in line with those introduced in Section 2.6.5, confirming that the integrated architecture does not have a significant effect on precision. However, the effect of higher DCR in the low photon count regime – which is a consequence of a multiple detector arrangement – is significant, causing a drop in precision of ≈ 5 dB from ideal at a total photon count of 100. The results therefore not only highlight the importance of acquiring as many photons as possible in a short time period, but also a requirement to scale the number of enabled detectors (N_D) with throughput to minimise the effects of noise at lower count rates. Finally, a balance must be made between achieving improved precision and minimising loss of accuracy at higher photon throughputs as the latter begins to deteriorate significantly beyond $\mu = 5.0$, as shown in Figure 3.23b.

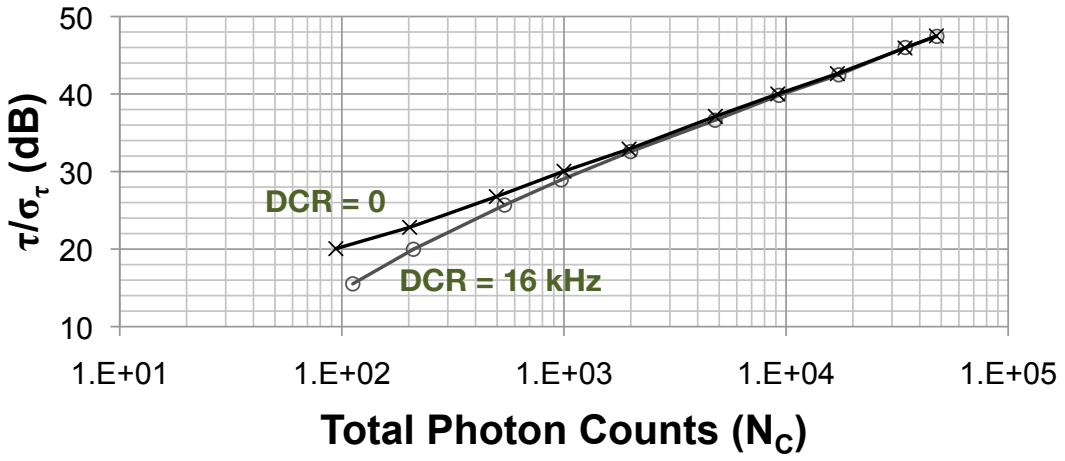


Figure 3.25: *Precision performance of CMM calculation within proposed integrated system architecture for varying DCR levels.*

3.10 Timer Mismatch

As described in Sections 2.2.4, 2.4.4 and 2.5.5, transistor mismatch will introduce gain errors between the timers in a multiple channel TI-TDC architecture. To study the effect of this gain mismatch on histograms and CMM calculations, the timer resolution variation (σ_T) and the number of TDCs required per TI-TDC timing channel to achieve zero dead-time (N_M) are now used for simulations. A 0.9 % (0.45 ps) standard deviation of the resolution has been reported for an array of 512 of the same TDC to be used in this implementation [12], so this will be used as the value for σ_T. For the purposes of this investigation, the TDC dead-time is assumed to be $\approx$ 10 ns (based on stand-alone circuit simulation trials) and the operating frequency is assumed to be $\approx$ 40 MHz (which is a typical laser diode frequency [136]), therefore from Section 2.5.5: $N_M = \lceil \frac{t_D}{f_s} \rceil + 1 = 2$. Furthermore, the model will now use *rev* to reflect the reversed start-stop mode (that will be used to minimise power consumption in the device) to highlight the issues caused by TDC gain mismatch. Finally, the chosen parameters of $N_T = 8$ and $N_D = 16$ are used together with $t_P/\tau = 0.01$ and $t_D/\tau = 0.4$ for all results presented in this section.

The design of the event distribution routing circuitry will have a significant impact on the resulting histograms and CMM calculations when gain errors are included. There are two primary techniques to distribute events to the timing channels: firstly, the router can be *reset* at the beginning of each excitation period, meaning the n^{th} photon arrival within each excitation period is routed to the n^{th} TI-TDC channel for processing; and secondly, the router can be *free-running* so that it always distributes each photon event to the next available TI-TDC channel, wrapping around to the beginning when necessary regardless of the excitation synchronisation (e.g. after every 8^{th} photon event in this case).

The graphs in Figure 3.26 show the contribution of each TDC (red/blue) to the total histogram (black) for the *resetting* router (top) and the *free-running* router (bottom) and for a photon rate $\mu = 1.0$. The resolutions of each TDC used to produce both graphs are shown in Table 3.3. In the case of the *resetting* router, the mismatch of the TDCs in the first timing channel dominate, producing a severe distortion at the histogram peak. In both cases, the TDC mismatch introduces an addition to the instrument response function (IRF), widening it proportional to the mismatch. The IRF increase in the *free-running* router approach is solely dependent on the mismatch, as each TDC has an equal utilisation. However, the utilisation of timing channels in the *resetting* router approach is very dependent on the photon rate, so the IRF is also very photon rate dependent.

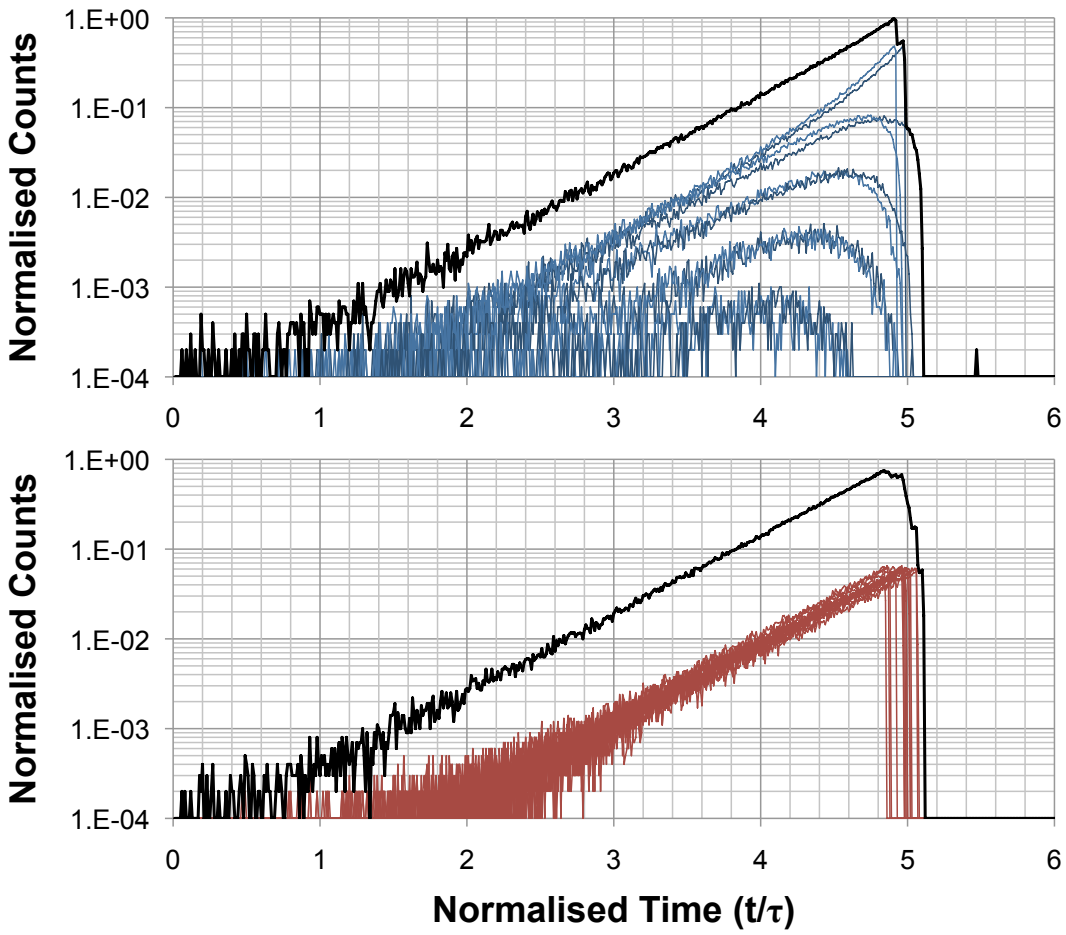


Figure 3.26: *Effect of TDC mismatch on the captured histograms for resetting (top) and free-running (bottom) routers and a photon rate* $\mu = 1.0$.

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
50.24	48.98	50.14	49.80	51.61	49.39	50.33	50.32
50.83	50.39	49.41	50.15	51.25	51.37	49.97	49.91

Table 3.3: *Example TDC resolutions (ps) for 2×8 TDC array with 0.45 ps standard deviation.*

CMM calculation results are shown in Figure 3.27 for the *resetting* (top) and *free-running* (bottom) router approaches. Each of ten coloured (red/blue) curves in the graphs represents different (Gaussian) randomly generated TDC resolutions and the black curve is the ideal case for this configuration with no mismatch, as presented in Figure 3.23b ($t_P/\tau = 0.01$). The value for *LAST* (see Section 2.6.4) in the *free-running* approach can be found by calculating CMM (which computes the average) of the IRF at any reasonable photon rate. However, as the IRF of the *resetting* approach is photon rate dependent, choosing *LAST* is non-trivial. For this experiment, it is calculated at a fixed photon rate of $\mu = 1.0$.

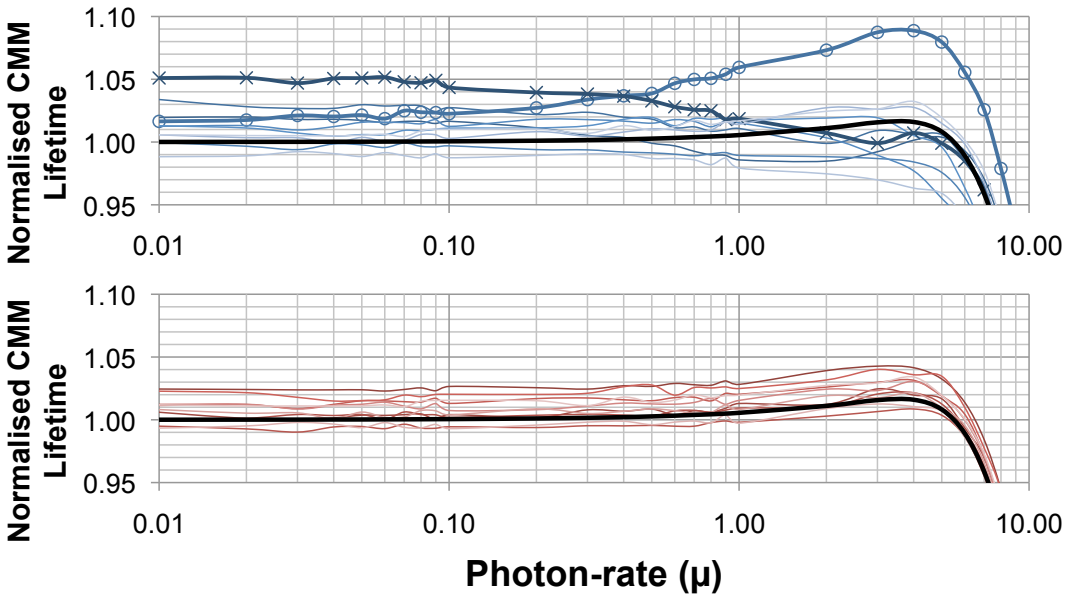


Figure 3.27: *Effect of increasing μ on the CMM calculation for ten different random TDC mismatch configurations using resetting (top) and free-running (bottom) routers.*

Performing this experiment with 100 random TDC mismatch configurations gives worst case errors of +10 / -3 % and +3 / -1 % for *resetting* and *free-running* router approaches, respectively, as shown in Figure A.4 in Appendix A. Furthermore, the error in the *free-running* approach is relatively constant with photon rate, so calibration is possible using known lifetimes. Conversely, due to the difficulty of choosing *LAST*, the error using the *resetting* approach has a much larger dependency on the photon rate, as shown by the two results marked with $\times$ and $\circ$. Due to the ease of choosing *LAST* and the possibility to calibrate, the *free-running* approach, which provides 3 times improved error performance, is clearly most suitable for the implementation of this architecture.

3.11 Conclusions

This chapter has introduced a MATLAB model and simulation environment that is used to investigate the different design parameters of the high photon throughput TCSPC architecture chosen from Chapter 2. The investigations look at the effect of each parameter individually on captured TCSPC histograms and fluorescence lifetime calculation using CMM, before combining them to provide an informed decision on their selection for implementation. The proposed parameters are used to highlight the expected device performance and limitations, before the chapter finishes by investigating the effect of TDC mismatch.

The chosen controllable parameters of the architecture – $N_T = 8$ and $N_D = 16$ – are shown to enable maximum photon throughputs in excess of $\mu_{max} = 4$ and $\mu_{max} = 10$ for 1 % and 5 % errors in lifetime calculation, respectively. This is an improvement of two orders of magnitude over an equivalent single channel TCSPC system, which is only capable of $\mu_{max} = 0.04$ for a 1 % error. For 5 % error, a photon throughput improvement of over 50 times is possible. The maximum photon rate is shown to be dependent on t_P/τ, however it has been demonstrated to provide *at least* one order of magnitude improvement for $t_P/\tau < 0.2$. For $t_P/\tau \leq 0.1$ a maximum photon rate in excess of the excitation frequency is also shown to be possible.

As discussed above, the maximum available photon rate is still very dependent on the channel pulse-width to lifetime ratio (t_P/τ) – and to a lesser extent the detector dead-time to lifetime ratio (t_D/τ). Assuming a SiPM output pulse-width of 500 ps, the architecture proposals provide a throughput in excess of the excitation rate for $\tau \geq 5$ ns. This highlights the architecture's suitability for measuring longer lifetime fluorophores, which are most limited by *classical* pile-up due to the extended excitation period required to resolve the decay. Reducing the SiPM output pulse width (t_P) as much as possible is clearly advantageous to allow better performance from shorter lifetimes, however it cannot be reduced indefinitely as it will still be limited by process and design constraints.

The inclusion of TDC mismatch in the model and simulation is shown to provide an additional source of error in the calculation of the fluorescence lifetime by CMM. This additional error is minimised by using a *free-running* event distribution router to give each TDC equal utilisation. This is made possible by simplifying the choice of *LAST* in the CMM calculation by capturing CMM of the IRF. Furthermore, the lifetime error is almost constant with photon rate, so can easily be corrected for by calibrating the sensor with fluorophore(s) of known lifetime(s).

All of the knowledge gained from the investigations throughout this chapter will be used to aid the design and verification of a high photon throughput fluorescence lifetime sensor. This is particularly important with regards to the design of the SiPM output pulse width (t_P), which has proved to be the limiting factor in this design, and will be presented in Sections 4.3.3 and 4.3.4. It is also important to understand the effect TDC mismatch has for the design and verification of the event distribution, which is presented in Section 4.4.3. Furthermore, the expected performance of the proposed parameters, introduced in Section 3.8.3, will provide a comparison when analysing results from the manufactured sensor in Chapter 5.

4
HIGH THROUGHPUT FLUORESCENCE LIFETIME SENSOR

4.1 Introduction

4.1.1 Background

Using the proposed architectural specifications from Chapter 3, a test chip was designed and manufactured to demonstrate the feasibility of overcoming the pile-up limit in fluorescence lifetime experimentation. Thanks to an ongoing working agreement between The University of Edinburgh and industrial partner STMicroelectronics, access was available on their 130 nm imaging process as part of a multiple project wafer (MPW). High performance SPAD structures [59] and TDC architectures [12] have recently been developed in this process, as introduced in Sections 2.5.2 and 2.5.4, respectively. This chapter describes the architecture of the sensor, *SIPM_CMM*, and the design and verification of the major blocks within the device. A photo-micrograph of the manufactured sensor is shown in Figure 4.1.

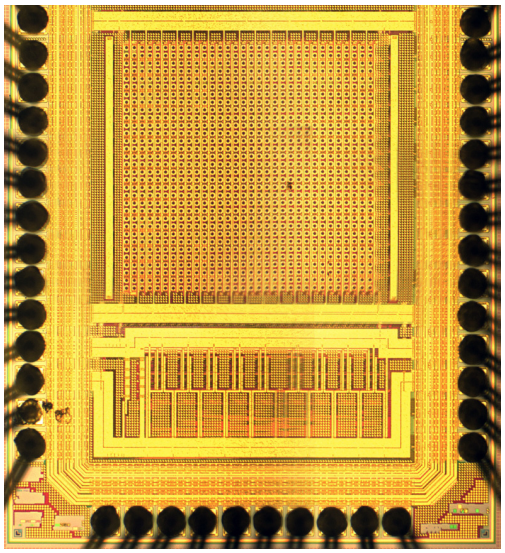


Figure 4.1: *Micrograph of SIPM_CMM test chip fabricated in STMicroelectronics 130 nm imaging process, measuring 1.5×1.3 mm.*

4.1.2 Specification and Requirements

The primary goal for the design of a custom architecture for fluorescence lifetime sensing is to overcome the severe pile-up limit enforced by commercially available TCSPC systems. Chapter 3 has shown that photon throughputs at least an order of magnitude higher than these current upper limits are possible. This is achieved by using multiple detectors ($N_D = 16$), increasing the SiPM rate by reducing its output pulse-width, increasing the number of timing elements available per excitation period ($N_T = 8$) and removing processing dead-time using TI-TDCs.

The integration of all elements in the detection path of a fluorescence lifetime microscopy experimental set-up (detection, timing, processing and delay) is to be achieved, miniaturising the components to fit within a total silicon area of ≤ 2 mm^2. The fill factor of the integrated single photon sensitive detector should be at least 10 %, a low-risk target that is an order of magnitude improvement over in-pixel TDCs [12, 122]. The embedded synchronisation delay should have a range of at least ≈ 100 ns in ≈ 1 ns steps, to allow it to span a 10 MHz excitation period.

The full range of the TDC should be > 1 µs to allow the measurement of longer lifetime fluorophores for applications such as Oxygen concentration sensing [16, 38]. Additionally, for multiple detector experimental setups [76, 77], it is advantageous to allow the devices to be networkable as slave devices to reduce the I/O and control overhead that they require. Detailed specifications can be seen in Table 4.1.

Specification	**Value**
SiPM output pulse-width (t_P)	< 500 ps
# TI-TDC timing channels (N_T)	8
TDC Resolution (fixed)	≈ 50 ps
TDC Full Range	> 1 µs
Delay-line Resolution	≈ 1 ns
Delay-line Full Range	≈ 100 ns
Detector Fill-Factor	> 10 %
# Detector Elements (N_D)	≥ 16
Power Consumption	< 10 mW
PVT Calibration	External
# Networkable Devices	≥ 100
Device area	≤ 2 mm^2
# I/O and power pads	≈ 30

Table 4.1: *Device specifications.*

Finally, the device must be easily tested, characterised and calibrated, particularly in terms of detection and timing performance, as well as its ability to perform accurate fluorescence lifetime calculations. As both a test and a fully functional operational mode, the device should be configurable as a TCSPC sensor with a single timing element providing raw TCSPC data for standard post-experiment fluorescence lifetime data analysis.

4.1.3 IP Reuse

The primary focus of this thesis is the design, implementation and proof of operation of a high-throughput miniaturised fluorescence lifetime sensor. As introduced in Section 1.4, both the single photon avalanche diode (SPAD) structure and time-to-digital converter (TDC) architecture are re-used intellectual property (IP) blocks from previous and current research projects developed in the same 130 nm imaging process.

The high performance blue sensitive SPAD [59], that is described in detail in Section 2.5.2, has been developed by Drs. Robert Henderson and Justin Richardson in collaboration with STMicroelectronics. The SPAD used is a larger version (8 μm) of the detector that was incorporated in [22, 59, 122].

The gated ring oscillator (GRO) based TDC [12], introduced in Section 2.5.4, forms the basic building block for the timing circuitry in the *SIPM_CMM* device. The $\approx$ 50 ps resolution TDC has been proven to perform TCSPC based fluorescence lifetime experiments thanks to results from [10, 11, 122, 133]. Furthermore, as well as occupying a small area footprint (50×50 μm^2), the TDC is also capable of providing timing information in *real-time* with no latency.

4.2 Sensor Architecture Overview

As discussed in Chapters 2 & 3, a multiple element, single output detector with multiple timing channel architecture is to be implemented to perform high throughput fluorescence lifetime sensing. A digital silicon photomultiplier (SiPM) architecture [107] with per-pixel compression [114] is employed to perform the single photon detection. The SiPM is re-configurable, allowing individual SPAD detectors to be turned on or off independently as required by system and experimental constraints. The single output of the SiPM detector is distributed to an array of 16 TDCs using a token-passing event distribution circuit. During every excitation period,

half of the TDCs are available to process photon events whilst the other half are reset, ensuring TI-TDC operation with $N_M = 2$. The data produced from the array of TDCs is much too great to transfer off-chip, so a pre-calculation of a fluorescence lifetime estimation using the centre-of-mass method (CMM) [10, 134] is embedded on-chip making use of the data provided by all of the 16 TDCs.

A user configurable delay line is included to allow the excitation synchronisation signal to be positioned within the TDC timing window, as well as to allow the SiPM to be rapidly gated on and off to ensure that SPADs are active when they are mostly likely to receive fluorescence (i.e. not background or excitation) photons. To facilitate the networking requirement, a custom designed shift register serial interface is used to write to registers on and to read data off the chip. Using this technique, devices can be *daisy-chained* together to reduce the system overheads for multiple detector experiments. Additionally, the device can be configured in a number of test, calibration and characterisation modes, most significantly it can operate in a standard TCSPC mode with a *single* timing element and raw TCSPC output. A top level block diagram of the architecture showing the minimum I/O requirements can be seen in Figure 4.2.

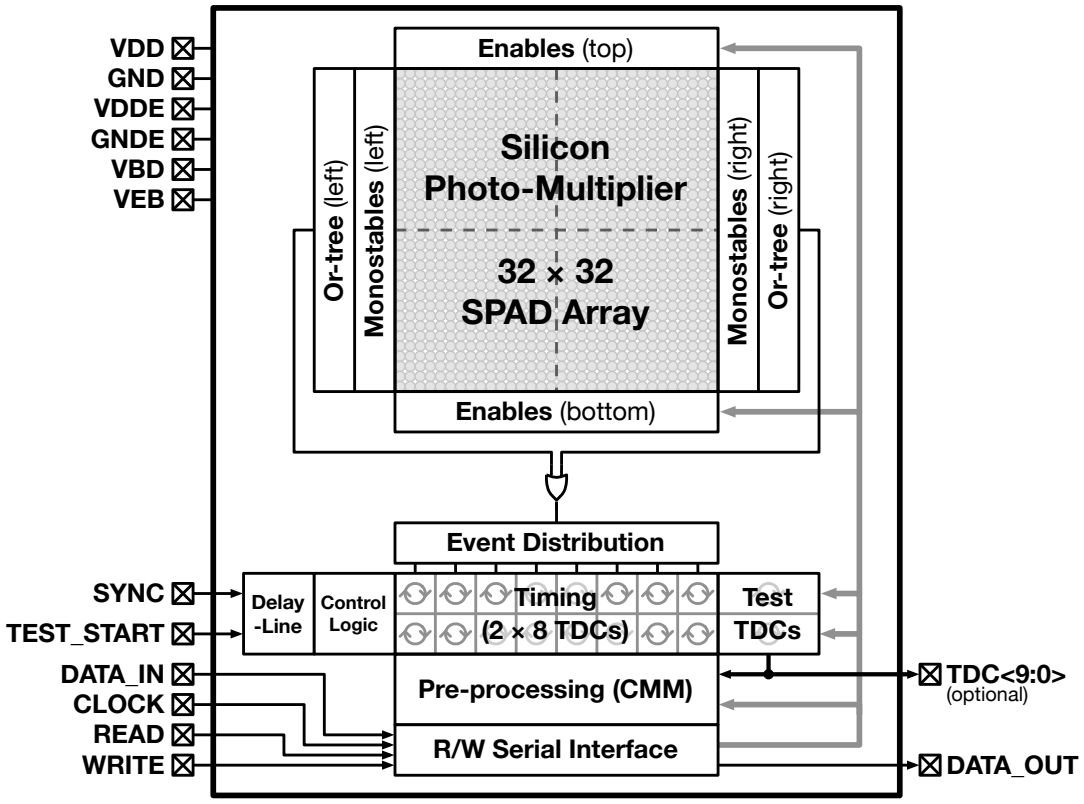


Figure 4.2: *System top-level block diagram.*

4.3 Silicon Photomultiplier

4.3.1 Overview

The digital SiPM design comprises 1024 circular, 8 µm diameter active area, negatively biased, passively quenched SPADs from [12], arranged in a 32×32 array. This large number of detectors was chosen due to the available silicon area and to provide experimental flexibility. The chosen value of $N_D = 16$ can be configured as each SPAD can be individually and independently turned on or off using static enable signals. To increase the maximum available throughput of the SiPM, the buffered output from each individual SPAD detector is passed through a pulse-shortening monostable circuit before being sent through a balanced OR-tree.

The SiPM is partitioned with the SPAD, quench and output buffer located inside the array, as shown in Figure 4.3. The monostable, OR-tree and enable circuits are then located at the periphery of the array as shown in Figure 4.2. This partitioning was chosen to maximise the fill-factor of the detector whilst keeping circuitry that is critical to the timing performance local to the SPAD. The timing critical components include the output buffer, which must drive long, high capacitance tracks and the SPAD quenching element which controls the detector dead-time. The resulting pitch of the SPADs and corresponding circuitry in the array is 21.5 µm. Enabling groups of adjacent SPADs provides a maximum active area of ≈ 0.05 mm^2, or a fill factor of just over 10 %.

4.3.2 Pixel

The circuit that is embedded within the SPAD array is shown in Figure 4.3 and contains the SPAD passive quenching element (p1), output buffer (NOR gate) and gating circuitry (OR gate and n1). The SPAD can be disabled locally using $\overline{ENABLE}$ or globally using *GATE* by disconnecting the Quench transistor and grounding the SPAD Cathode using n1. Additionally, the output must be pulled low as gating the SPAD off will cause the output to appear *on* and lock the OR-tree output. The functionality of the *GATE* and $\overline{FQUENCH}$ signals will be described in Section 4.7. Detector sensitivity is controlled using a variable high-voltage excess bias, V_{EB} (2.8 – 3.3 V), to power the SPAD, requiring thick-oxide transistors within the array. This power supply is controlled independently of the core V_{DD} (1.2 V) supply, protecting the timing and processing circuitry from light dependent power consumption. However, level-shifters are required so that inputs and outputs of the pixel are compatible with the core V_{DD} supply.

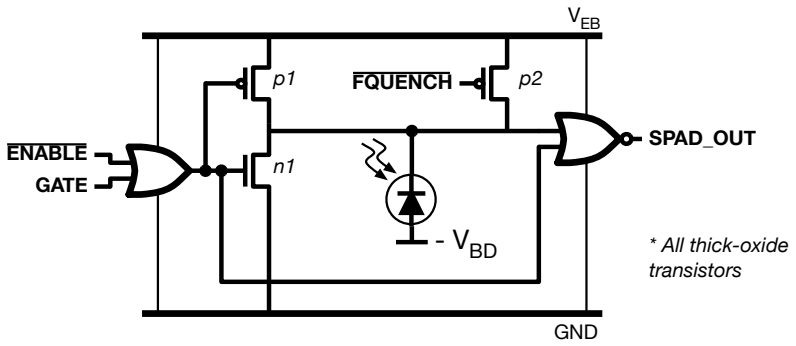


Figure 4.3: *Embedded SPAD quenching and output buffer circuit.*

As shown in Figure 4.4, the pixel circuit elements are positioned so as to maximise the fill factor of the SiPM. Alternate rows of detectors are mirrored vertically so that their accompanying transistors and standard cells sit physically beside one-another within the space between four SPADs. Furthermore, routing and power channels are formed over the metal1 (purple) guard ring of the SPADs, both vertically and horizontally. The SPAD outputs are distributed horizontally on metal2 (turquoise) from the centre towards the left and right of the array. The individual enable signals for each SPAD are distributed vertically on metal3 (pink) from the top and bottom to the centre of the array. In both of these cases, a routing channel 16 wires wide is required to connect each half of the 32×32 array. The three power supplies, V_{EB}, V_{BD} ($\approx$ –13 V) and *GND* are then distributed in a grid in the remaining space.

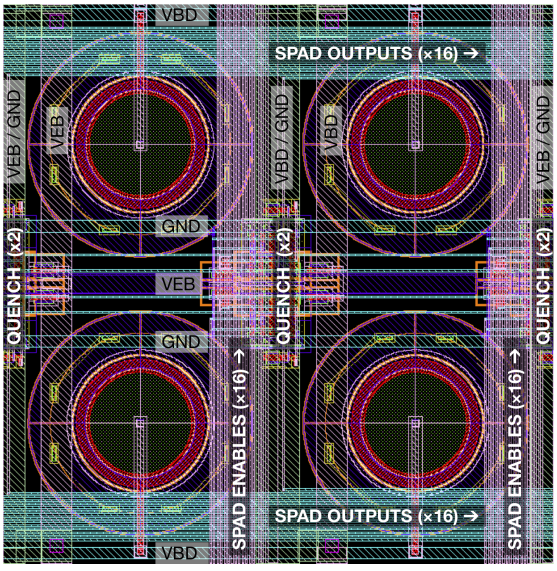


Figure 4.4: *Annotated layout of 2×2 SPADs from the bottom right-hand corner of the SiPM.*

4.3.3 Pulse Shortening

To implement the pulse-shortening, which prevents the dead-time of individual detectors in the SiPM from restricting the maximum count rate, the buffered output from each individual SPAD detector is passed through a monostable circuit, as first introduced in Section 2.5.3 and shown conceptually in Figure 4.5.

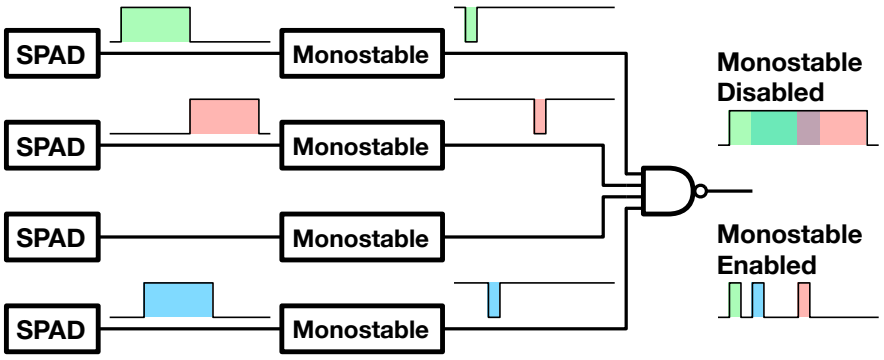


Figure 4.5: *SPAD output compression concept.*

The monostable circuit, as shown in Figure 4.6, is implemented by NORing (I5) the SPAD pulse with a delayed version of itself (delay through I2+I3+I4), where the pulse width at the output is equal to the total delay time. The circuit has an inverter with thick-oxide transistors (I1) powered by V_{DD}, whose input is overdriven, to level shift the SPAD output back to the core supply voltage. Additionally, all of the pulse-shortening circuits can be enabled or disabled using the global $\overline{MSENABLE}$ signal driving I4. Although the output of the monostable circuits will have short pulse widths, the individual SPADs that created these events will remain insensitive to subsequent photon events within their own inherent dead-time. This creates a spatial *pile-up*, which is minimised by the use of many small active area detectors within the SiPM architecture, as investigated in detail in Chapter 3.

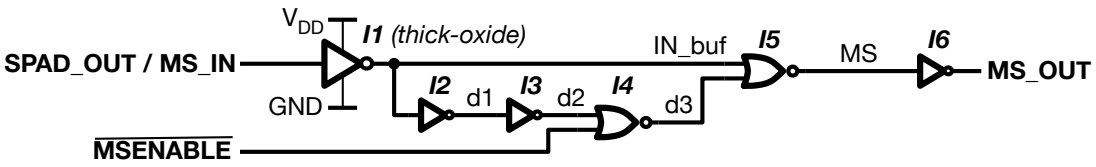


Figure 4.6: *SPAD output pulse-shortening monostable circuit.*

As Chapter 3 concludes, the width of the pulse should be as short as the process and architecture constraints allow to keep the likelihood of detector pulse-width based pile-up to an absolute minimum. Waveforms from SPICE simulations of the monostable circuit, as shown in Figure 4.7, show the minimum rail-to-rail pulse-width, limited by the minimum gate delay of the process, to be ≈ 40 ps, using medium buffer strength *High-Speed* logic for all standard cells.

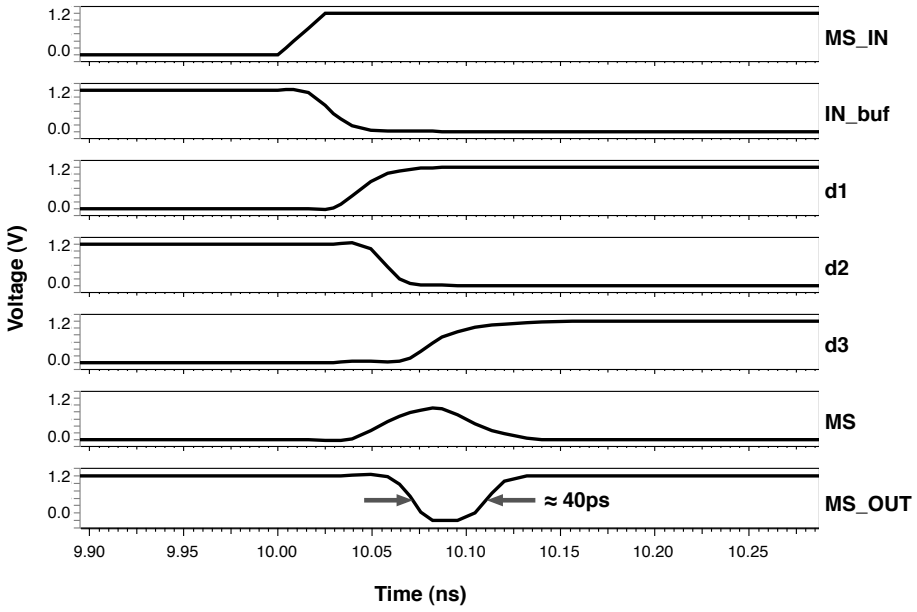


Figure 4.7: *Simulation of process limited monostable output at 40 ps.*

4.3.4 OR-Tree

An OR-tree is required to combine the outputs from individual SPAD elements into a single detector output for the timing circuitry. To minimise delay for the timing critical TCSPC application, the OR-tree is implemented in negative logic using *High-Speed* standard cells with increasing buffer strength to compensate for the increasing length of track, and therefore increasing capacitance, that each stage must drive. This is shown in Figure 4.8.

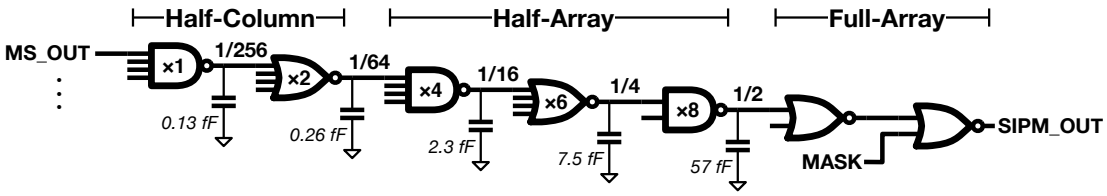


Figure 4.8: *Or-Tree Schematic showing increasing buffer strength for a single detector path.*

The increasing track length also limits the monostable pulse-width as it must be long enough to safely pass through the tree and trigger the timing circuitry at the output. After thorough extracted SPICE simulations of the worst case pulse path through the OR-tree, as shown in Figure 4.9, a monostable output pulse-width of $\approx$ 250 ps was chosen. This is four times shorter than [114] and two times shorter than the estimated value from Sections 3.7-3.11, enabling an improvement in throughput performance for shorter lifetimes. The pulse-width is created by increasing the minimum transistor gate width and length of the NMOS and PMOS in the custom delay inverter (I3) in Figure 4.6.

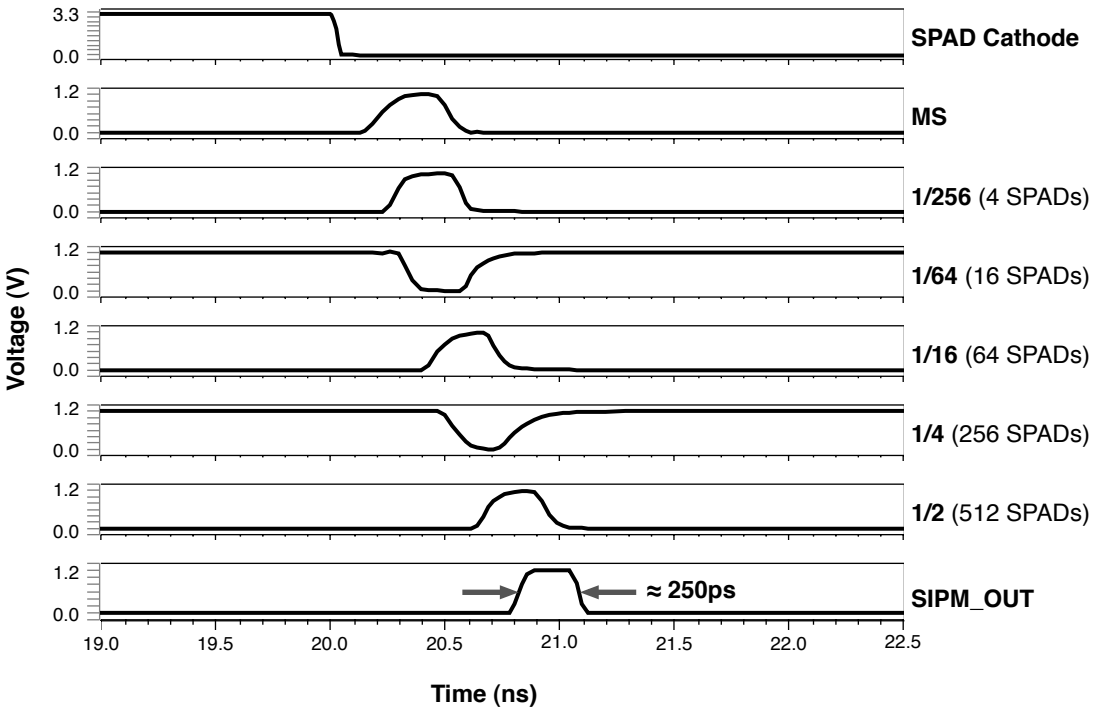


Figure 4.9: *Extracted simulation timing diagram detailing the worst case propagation of a shortened SPAD pulse through the OR-Tree.*

4.3.5 Enables

Each SPAD in the SiPM can be individually and independently turned on or off using static enable signals, as described in Section 2.5.3. This not only provides a method for monitoring each SPAD individually to measure performance characteristics such as dark count rate (DCR) and timing, but also acts as a method to enable a sub-array for experimentation. Noisy detectors that would otherwise have a negative effect on the signal-to-noise ratio (SNR) can also be disabled, at the expense of reduced sensitivity.

The enable signals are implemented using an extension of the read/write serial interface as described in Section 4.6.1. In short, it consists of a 1024-bit chain of D-Type flip-flops arranged as a shift-register that hold the state of each SPAD in the SiPM. The flip-flops are designed in standard core logic, so a level-shifter from V_{DD} to V_{EB} is required at the output of each enable signal to allow it to drive the in-pixel circuitry. As with the monostable and OR-tree circuitry, the enable shift-register and level-shifters are located outside the active area of the SiPM.

4.4 Multiple Channel Timing

4.4.1 TDC

The timing element of the multiple channel timing architecture is the $\approx$ 50 ps time-to-digital converter (TDC) from [12] that was developed in the same 130 nm process and used in [22]. It was chosen for its small area footprint and ability to provide timing information in *real-time* with no latency. A block diagram of this TDC is shown in Figure 4.10. Although the gated ring oscillator (GRO) core of the TDC, as described in Section 2.5.4, is a reused IP block, minor modifications were necessary to its front-end logic (red) to make it function correctly within the proposed multi-channel timing architecture, rather than the image sensor read-out for which it was originally designed.

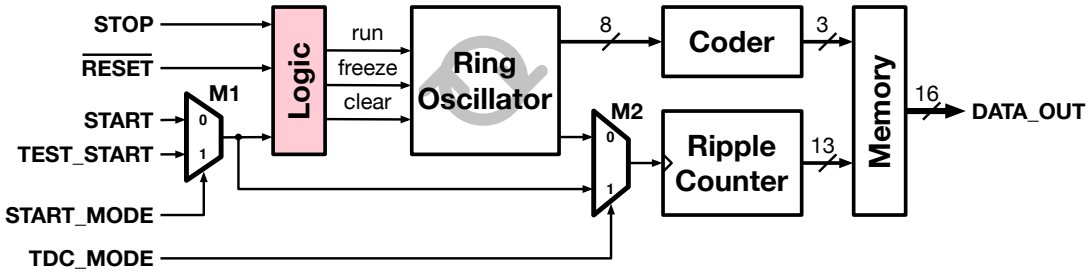


Figure 4.10: *Time-to-digital converter (TDC) structure, modified from [59].*

As the TDC is no longer area constrained due to pixel dimension limitations, it was possible to increase the size of the ripple-counter from 7 to 13 bits. This has the effect of significantly increasing the full range of the TDC from 50 ns to over 3 µs, meeting the $>$ 1 µs specification to allow longer lifetime fluorophores to be measured for applications such Oxygen concentration sensing. The original pixel based TDC contained the SPAD detector, so adding additional bits to the ripple-counter only increases the total area by approximately 10 % (to 44×62 µm).

The TDC will operate in a reverse start-stop mode (see Section 2.2.2), so that power consumption is proportional to photon activity. However when running, the TDC will draw $\approx$ 300 µA of current, which will be a high proportion of the total device power consumption. Therefore, the stability of its power supply is critical to keeping a fixed TDC resolution, independent of light levels and processing activity. For these reasons, the TDC has a power domain, V_{DDOSC}, separate from V_{DD} with substantial local decoupling. However, due to the low number of available pads, these power supplies are connected on-chip at the V_{DD} pad.

The TDC can also be configured to *count* SiPM events rather than time their arrival, by simply using the ripple-counter and configuring M2 accordingly to bypass the GRO. Together with the raw TCSPC read-out mode (as described in Section 4.8.1), this technique allows the sensor to be used in photon counting, or time uncorrelated single photon counting applications such as fluorescence correlation spectroscopy (FCS) [14]. Furthermore, the TDC can be configured to accept a *TEST_START* signal, in place of the SiPM output, by configuring M1 accordingly. As will be described in more detail in Section 4.8, this is critical to the test, characterisation and calibration of the device.

4.4.2 Time-Interleaved TDCs

As introduced in Section 2.3, one of the major limitations of single timing channel TCSPC systems is processing dead-time. For the TDC introduced above, this is the time required to *write* the current timestamp to memory and then *reset* the ring oscillator, and is shown to take less than 20 ns using circuit simulations. Assuming a maximum excitation frequency of 40 MHz [136], the number of converters in a time-interleaved architecture to remove this dead-time is: $N_M = \lceil \frac{t_D}{f_s} \rceil + 1 = 2$ (see Section 2.5.5). Therefore the TDCs will operate as time-interleaved *pairs*, so that one TDC is active and available to accept a single photon event while the other is dead. The TDCs are entirely controlled by the excitation synchronisation pulse, which is typically a short impulse, so is initially divided by two using a toggle flip-flop, as shown in Figure 4.11. This creates two out of phase clocks, whose positive going edges act as the *STOP* signal to the pair of TDCs, so each is only active during the low cycle of its *STOP* signal. The *READ*, *WRITE* and $\overline{RESET}$ signals for the TDCs are then generated using logical combinations of different tapped outputs from an inverting delay line that uses one of the *STOP* signals as its input. The TDC time-stamps are written to a local memory before being read-out via a shared 16-bit bus on the following excitation cycle.

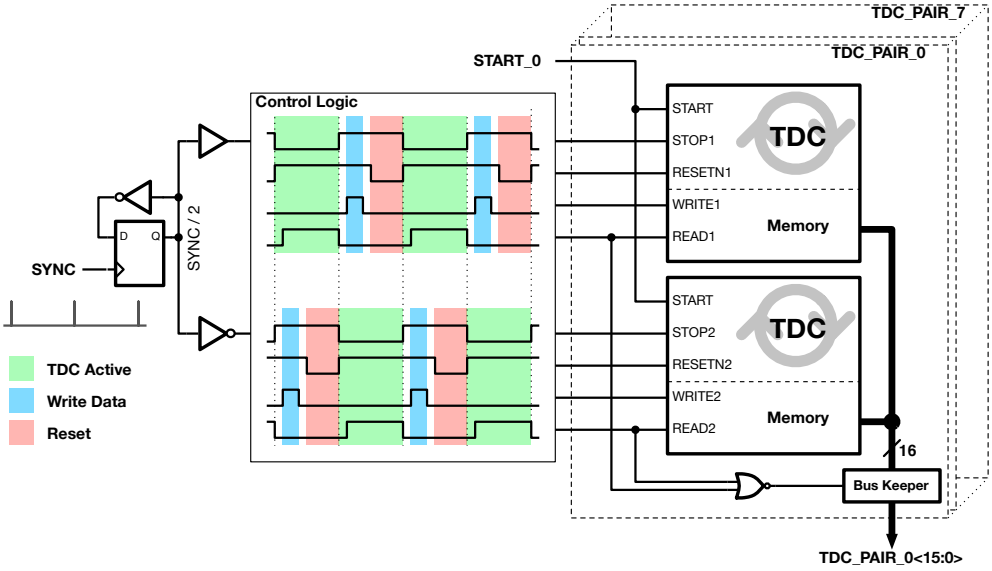


Figure 4.11: *Timing generation for interleaved TDC pairs.*

SPICE simulations of this architecture were carried out using a 20 MHz excitation pulse and an asynchronous but periodic 15 MHz event pulse as stimulation. This simulation provides a TDC resolution of $\approx$ 33 ps, which is just over 50 % faster than what is extracted or measured on the bench. The timing diagram in Figure 4.12 shows the operation of the major signals in the TDC-pair architecture, highlighting in red and blue photon events being timed, stopped, written and read out by the first and second TDCs, respectively. The output is pipelined so that the TDCs can remain active whilst reading data out and is updated on every excitation pulse, with a single cycle latency.

The photon event highlighted in green in Figure 4.12 underlines a special case of operation for this circuit, where the SiPM output transitions high just before the synchronisation signal and then remains high into the next synchronisation period. The control logic within the TDC itself has been modified to ensure that a SiPM event is only recorded within a synchronisation period if it sees a positive going edge. In this example, a TDC code of 2 is recorded for the positive going edge and the subsequent excitation period registers no events (code 0).

By running the simulation with the photon event stimulus delayed by a single excitation period, of 50 ns in this case, it is possible to simulate any timing mismatch between the two TDCs caused by non-idealities in the logic that creates the *STOP* signals. It was found that there was $\leq$ 1 TDC code difference between the TDC output for the same SiPM event to excitation time.

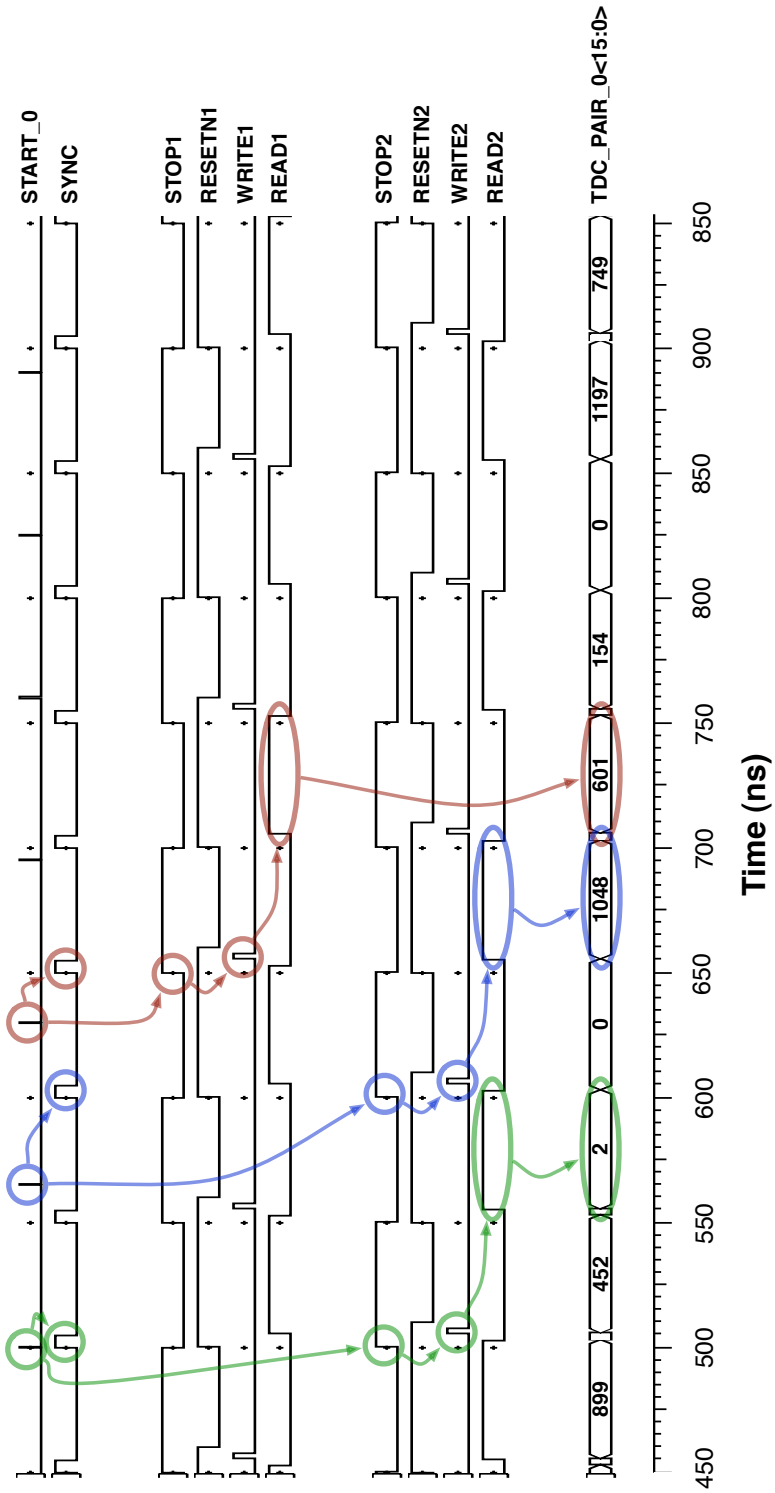


Figure 4.12: *Simulation of TDC pair timing showing SiPM and* SYNC *inputs, TDC control signals and TDC output data.*

4.4.3 Event Distribution to Multiple Timing Channels

A routing circuit is required to distribute arriving photon events to the proposed eight TI-TDC pairs ($N_T = 8$) from Section 3.8. The routing can be achieved using either a *resetting* or a *free-running* approach, with the latter being the preferred option to minimise TDC mismatch errors as described in Section 3.10. A free-running token-passing *ring* is the ideal circuit to implement this functionality, however simulation and bench testing show that the high-speed asynchronous nature of the photon event arrivals significantly increases the likelihood of flip-flop metastability in such a circuit, so the *token* can become corrupted or lost completely.

Therefore, a pair of interleaved token-passing *shift registers*, as shown in Fig. 4.13, is used to distribute events to the array of TDCs. The token-passing shift register circuits operate in a similar way to the TDC-pairs, where they are active on alternate excitation cycles while the other is reset to prepare it for the next excitation cycle. The shift registers are 15 elements long and alternating bits are clocked using complementary edges of the SiPM output. This ensures that the next TDC-pair in the array does not have a photon event passed through to it until the previous event has completely finished (signified by its falling edge).

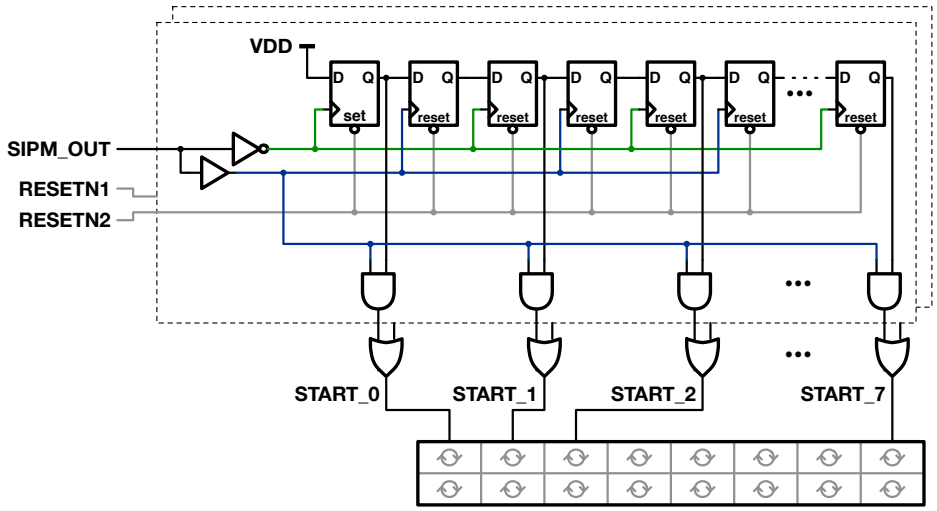


Figure 4.13: *Token-passing circuit for SPAD pulse distribution to TDC pairs.*

Metastability is still an issue in this circuit due to the high speed asynchronous event arrivals. To minimise this effect, the token is not cleared from the previous flip-flop in the chain, so that if a state goes metastable it will be corrected by the next photon event. Each TDC will only time the first event it sees, so not clearing the token from the previous register does not affect the operation of the architecture. Verification of this circuit will be included in Section 4.5.4.

4.5 Embedding the Centre-of-Mass Method

4.5.1 Overview

Embedding some form of fluorescence lifetime calculation on-chip was one of the original aims, both to miniaturise the standard experimental set-up and to remove the requirement of intense software post-processing of data. Furthermore, by removing the need to send data to a CPU to calculate lifetimes, we introduce the concept of providing lifetime calculations in *real-time*, which opens the way for new applications of time-resolved fluorescence lifetime sensing, such as flow cytometry (see Section 1.2.5).

Although *real-time* solutions have been performed by using the parallel computing capabilities of FPGAs [10], large amounts of data still needs to be distributed off chip at high rates. The proposed timing architecture creates up to 128-bits (16·8) per excitation cycle, or a data rate of $128 \cdot f$ Mbps (where f is the excitation frequency in MHz). Using an 80 MHz I/O data rate, such throughput would require $1.6 \cdot f$ parallel output pads, or 32 for a conservative 20 MHz excitation rate. As well as not being scalable, such an architecture would require a highly-parallel I/O that would significantly increase I/O power and chip area, creating a severely pad-limited design, one of the major drawbacks of [12, 122]. Therefore embedding some form of processing is the only realistic solution to keep chip area below 2 mm^2.

As introduced in Section 2.6, the chosen fluorescence lifetime calculation technique is the centre-of-mass method (CMM), selected for its use of digitised TCSPC data and its high photon efficiency (ideally 100 %). As shown by Equations 2.12 and 2.13, the core of the CMM calculation is the computation of the average TDC code. It is necessary therefore to accumulate the total ***sum*** of TDC codes and ***count*** the total number of photon events.

Power consumption and area constraints mean that full integer division on-chip is not possible. Accumulating a power of two photons allows the division to be performed using a binary shift. However, performing the division in this way, as implemented on-FPGA for [10, 134], has a major drawback; results are only updated after the power of two counts is reached, so the update rate is *photon-rate* dependent and the photon count information is lost. For these reasons, it was decided to not perform the division on-chip, but to transfer the total sum and total count to an external device such as an FPGA or microcontroller to perform the final division and preserve the count rate information. Furthermore, to ease debugging of the algorithm, the background correction will also be partitioned to this external device.

4.5.2 TDC Interface

The first stage of the CMM calculation is the summation and counting of events from each pair of TDCs, a block diagram of which is shown in Figure 4.14. The data from each set of two TDCs will be driven onto a shared 16-bit bus on alternating synchronisation, or clock cycles. In order to provide experimental flexibility, compensate for synchronisation offsets and improve SNR performance, only TDC codes that fall within a pre-defined measurement window are included in the calculation (see Section 2.6.4). To implement this, a digital comparator, synthesised from compiled Verilog, is placed on the 16-bit data bus, taking global register values *FIRST* and *LAST*, that define the window's position and width. Only data that falls inside this window between *FIRST* and *LAST* is passed through to the next stage of the calculation. The window is configured according to the theory developed in [10]. As well as passing data through, the comparator triggers a 25-bit ripple counter to count the number of valid events.

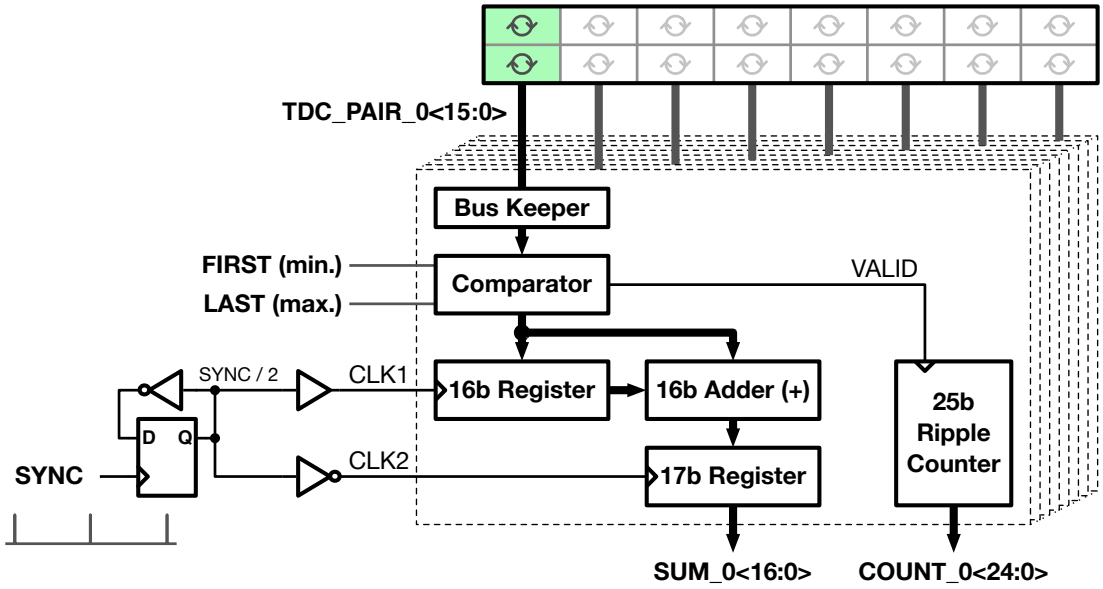


Figure 4.14: *First stage of the CMM calculation.*

The valid data from each TDC pair is then summed every two excitation cycles, implemented by registering the data from one TDC on one clock edge, then storing the result of an addition between it and the data from the second TDC on the inverse clock edge. The *CLK* signals are defined by the same trigger flip-flop as in Figure 4.11, as they have the same timing as the TDC *STOP* signals. The sum of each pair of TDCs is then produced on the edge of the second clock signal (*CLK2*).

4.5.3 Calculation

The final CMM pre-calculation is then performed using the data from all 8 sets of TDC pairs, as shown in Figure 4.15. Summation of valid time-stamps is performed using a pipelined adder tree, operating on alternating clock edges, followed by a 36-bit accumulator. The total number of valid TDC events is calculated using a combinatorial adder tree, providing a 28-bit result. Although the summation could also be performed using a combinatorial adder tree, by placing an accumulator at the output of each TDC pair, this would double the data bus widths, causing routing congestion on this optically optimised 3-metal process [97].

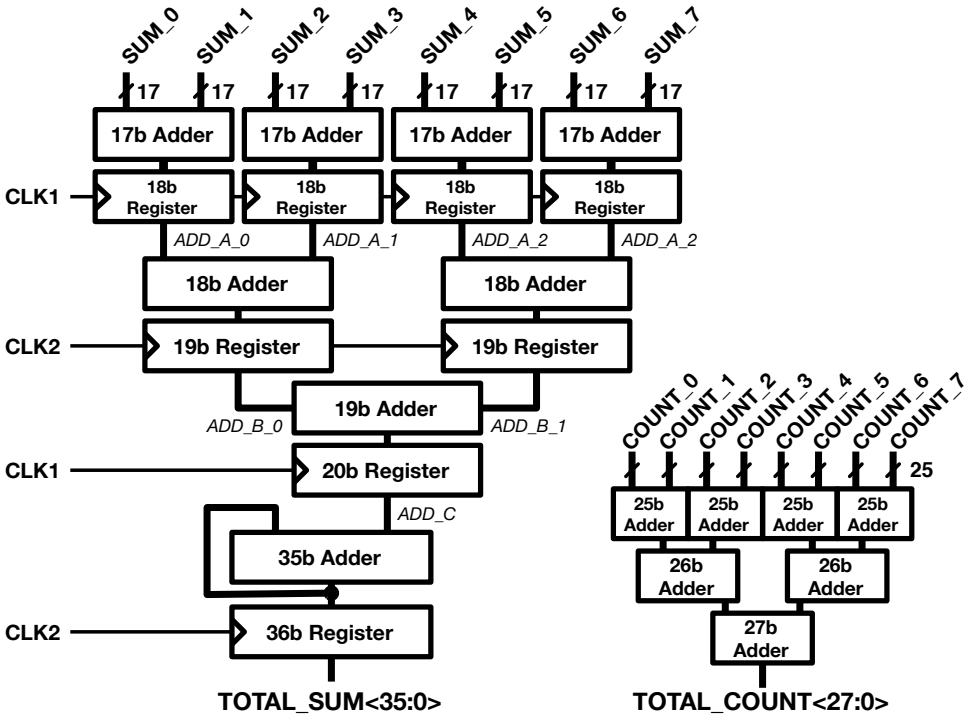


Figure 4.15: *Implementation of CMM pre-calculation.*

Alternating the clocks to the registers of the total sum calculation allows the final result to be produced twice as fast, in this case two clock cycles faster, than with a single clock source. When sampling the CMM data to be read off-chip using the serial interface, the inputs of the TDC architecture need to be gated to allow the data to settle. This introduces a read-out dead-time that the alternating clocking reduces by a factor of two. The total sum (36-bit) and count (28-bit) values are then sent off-chip periodically for further integration and used to calculate the final lifetime estimation, corrected for background noise and other calibrated non-idealities such as TDC resolution variation.

4.5.4 Functional Verification

A SPICE simulation of the entire processing system, including token-passing event distribution, TDC array and CMM pre-calculation is performed to verify the correct operation of the combined circuits. Results of CMM summation and count data, together with the intermediate steps of the calculation, are shown in Figure 4.16, where the system is stimulated with a 20 MHz excitation synchronisation pulse and 3 photon events per excitation. The photon events are asynchronous with each other and with the excitation to create a range of different TDC time-stamps for each event.

After being reset at the beginning of the simulation, the system is active for the first 10 excitation periods (just under 500 ns). The raw TDC time stamps produced by the first 3 sets of TDC pairs are shown by the signals *TDC_PAIR_**, the data of which is updated on every excitation period as explained in Section 4.4.2. The measured time stamps are consistent with the time delays between the stimulated photon events and the excitation *STOP* signals.

To test the windowing functionality of the CMM algorithm, the comparator at the output of the TDC pair has been setup to only pass through events, or time stamps, that fall within the window between *FIRST* = 96 and *LAST* = 320. The TDC codes are highlighted to show which fall within the window (green text) and those that do not (red text). The summation of consecutive valid time stamps for each TDC pair, as explained in Section 4.5.2, is then highlighted in green with the output signals being shown by *SUM_**.

Each (non-zero) stage in the pipelined adder tree, as described in Section 4.5.3, is highlighted in blue with the outputs of each stage shown propagating through by the signals *ADD_**. The final accumulator step is then highlighted in red, producing the final 36-bit summation of all TDC codes as shown by *TOTAL_SUM*. Finally, the total number of valid events is shown by the 28-bit value *TOTAL_COUNT*.

The total summation of 4,807 and total count of 25 is verified by manually adding and counting the valid raw TDC values produced by the TDCs. It can be clearly seen that the worst-case propagation delay is through the pipelined adder and accumulator to calculate the total sum of TDC events. A total time of six excitation periods is required before the output data can be sampled.

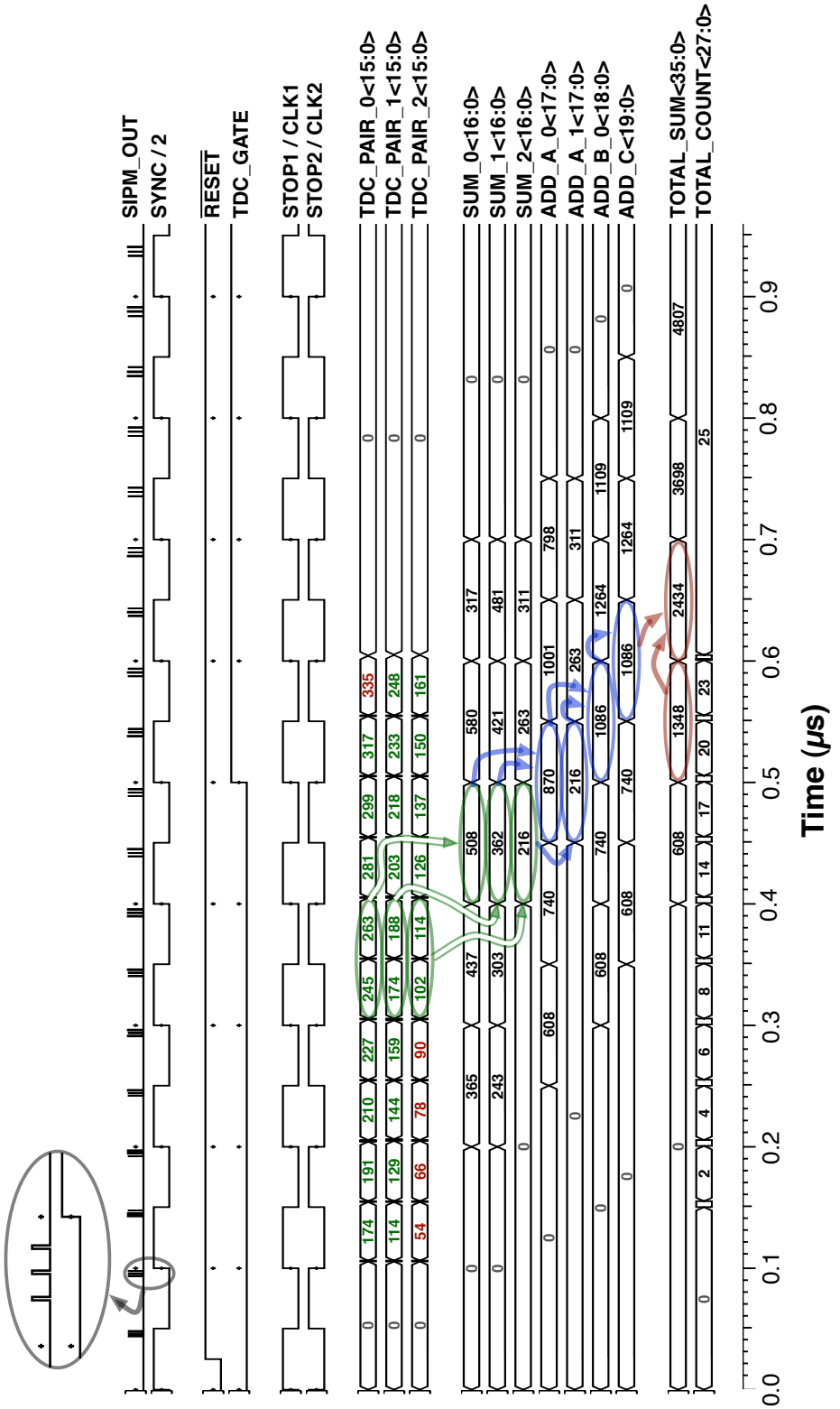


Figure 4.16: *Simulated results from TDC and CMM timing.*

4.6 Device Communication and Control

4.6.1 Custom Serial Interface

A bi-directional serial interface is required to allow efficient read access to CMM data as well as write access to on-chip registers and SPAD enable signals. It must be designed to meet the low pad count and small area specifications outlined in Section 4.1.2. Furthermore, the interface should allow multiple devices to be networked. A total of 64-bits are required for reading off CMM data (28+36 bits) and writing to the embedded register bank, details of which can be found in Appendix A.6.

The core of the design to meet these requirements is a 64-bit shift register, as shown at the top of Figure 4.17. A read operation is performed by holding *READ* high and strobing *CLK* once to load the data into the shift register, then by holding *READ* low and strobing *CLK* 64 times, data can be sampled off-chip using the *DATA_OUT* signal. A write operation is performed by setting *DATA_IN* as required and strobing *CLK* for each of the 64 bits, then the *WRITE* signal is strobed to store the data in a second bank of 64 flip-flops. With an 80 MHz *CLK* and *DATA* rate, a maximum readout rate of 1.25 / n MHz is possible (where n is the number of devices in the network), providing theoretical exposure times of $< n$ µs. However, with read-out dead-times in the region of hundreds of nanoseconds, longer exposure times are preferred.

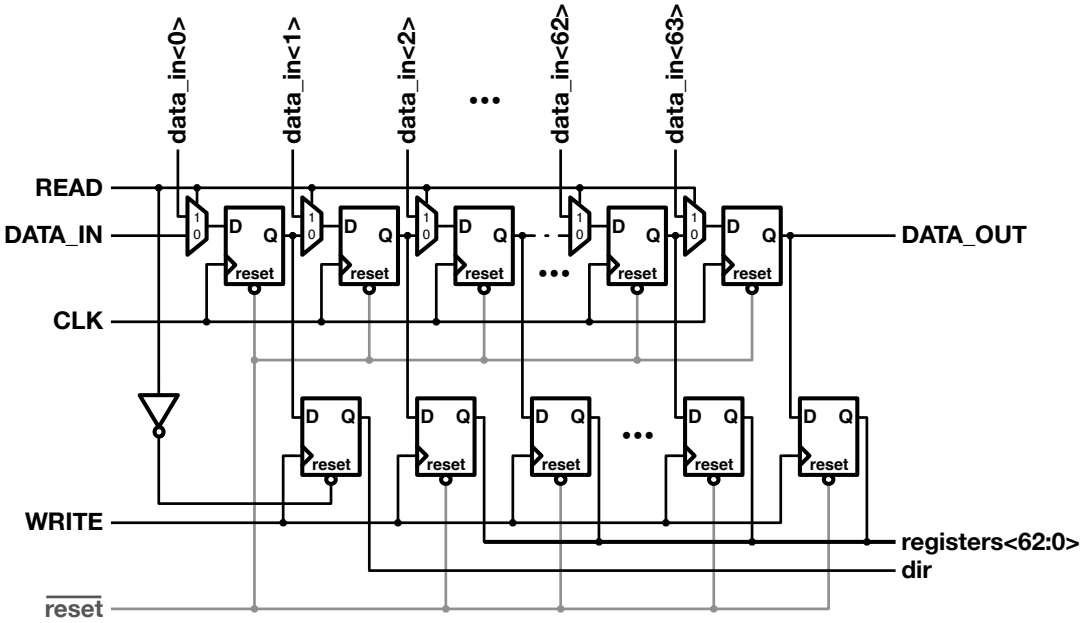


Figure 4.17: *Custom bi-directional shift-register based serial interface.*

The use of a shift register as the serial interface allows devices to be connected, or networked, in a *daisy-chain* configuration, taking inspiration from the *LT1446* DAC. To allow devices to be daisy-chained in close proximity, the *CLK*, *READ* and *WRITE* signals all need a corresponding **_OUT* signal to control the next chip down the chain. Additionally, the *CLK* signal is routed in the opposite direction to the *DATA* signal to negate the likelihood of flip-flop shoot-through.

It is also necessary to control the enable bits of the SPADs using the same serial interface. This could be achieved by adding 1024 to the length of the shift register, but doing so would require reads and writes to take 16 times longer. Therefore *DATA* and *CLK* signals have been designed to fork at their input depending on a control bit (*dir*) and recombine at the output, using a multiplexer controlled by the same bit, to facilitate the *daisy-chaining* functionality. The logic overhead to implement this is minimal and is shown in Figure 4.18, where the *LOGIC_LEFT* and *LOGIC_RIGHT* blocks correspond to the physical location of the logic on the device. The control bit (*dir*) can be set by performing a register write, temporarily providing access to the enable shift-register. Strobing the *READ* signal (unused for the enables) will reset *dir*, returning functionality to the serial-interface shift-register, as shown by the first memory element in Figure 4.17. It should be noted that buffering elements are not shown. The 1024-bit enables shift register does not require a memory or a *WRITE* signal as it does not have the same dual read/write functionality as the main 64-bit shift register.

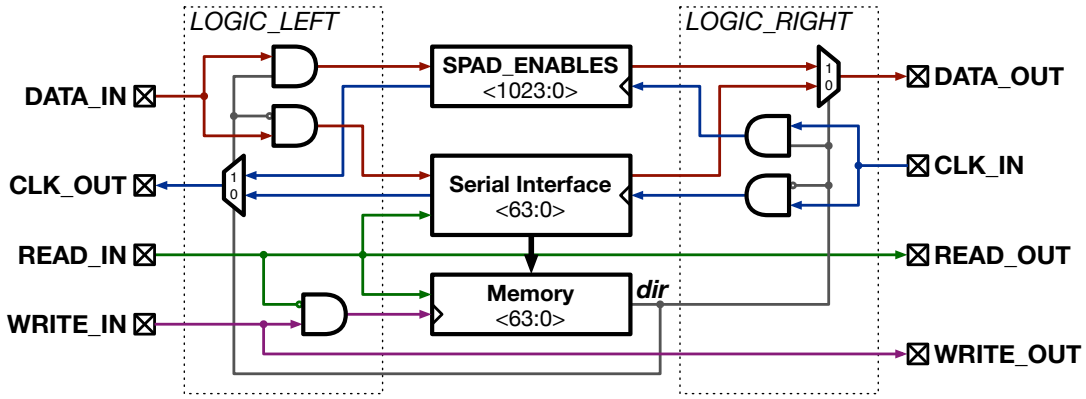


Figure 4.18: *Serial Interface control-logic blocks.*

To keep the number of pads to a minimum, specific combinations of the four serial interface input signals are used to perform system and block-level resets. A system reset is performed by strobing *READ* and *WRITE* high simultaneously. To stop the contents of the serial interface being written to the register memory during this time, the *WRITE* signal is gated when *READ* is high, as shown in Figure 4.18. Furthermore, the *READ* signal is used to gate the inputs to the TDC architecture, allowing data to settle at the output of the CMM calculation, as discussed in Section 4.5.3, before the *CLK* signal latches the data into the serial interface shift register. A delayed version of the logical AND of *CLK* and *READ* is then used to reset the TDCs and CMM blocks ready for the subsequent exposure. More information on the timing of these signals as well as the timing of serial-interface reads and writes can be found in Appendix A.7.

4.6.2 Networking

The ability to network devices in some way is important to reduce the system overheads for multiple detector experimental set-ups [76, 77]. Typical multi-detector set-ups require individual power supply and signal connections for each detector, routed to individual or multi-channel timing cards. As well as being a bulky and expensive approach, this technique does not scale well for tens of detectors.

As discussed in Section 4.6.1, devices will be connected using a *daisy-chaining* approach and controlled by a common set of signals. By using the shift-register serial-interface, devices can be networked by connecting the *DATA_OUT* of one device to the *DATA_IN* of the next device, as implemented with the *LTC1446* DAC and as shown in Figure 4.19.

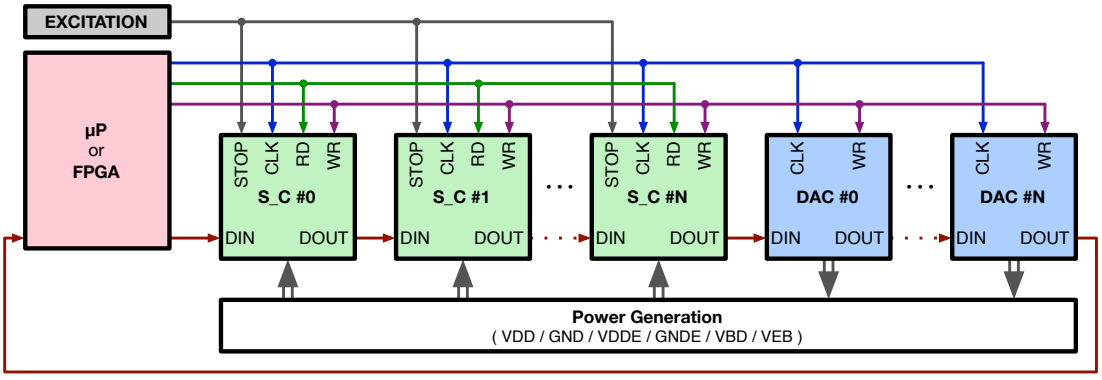


Figure 4.19: *Distributed daisy-chain network configuration.*

The additional serial interface signals: *CLK*, *READ* and *WRITE*, can then be connected as common signals to each device in the network. The technique requires data to be clocked through the serial interfaces of all devices for register operations, slowing down reads and writes. By using this approach, it is also possible to add *LTC1446* devices to the network, which are used to provide software controlled bias voltages for V_{EB} and V_{BD}.

However, this approach still requires the common control wires to be distributed individually from the controller to each device in the network, adding load to the signals. Furthermore, to network devices in close proximity, such as within the same package, not only do the *DATA_** signals need to be chained, but all other signals and power supplies need chained connections too. By re-buffering the control signals and re-enforcing power supplies on-chip, they can be sent *through* the devices in the network, as shown in Figure 4.20. Signals to be chained are located directly opposite each other in the pad ring, as is shown in Figure A.9. A total of 24 I/O pads are required for chaining functionality, consisting of 8 serial interface signals, 4 synchronisation signals and 12 power supply connections (more detailed information on the device pad list can be found in Appendix A.8).

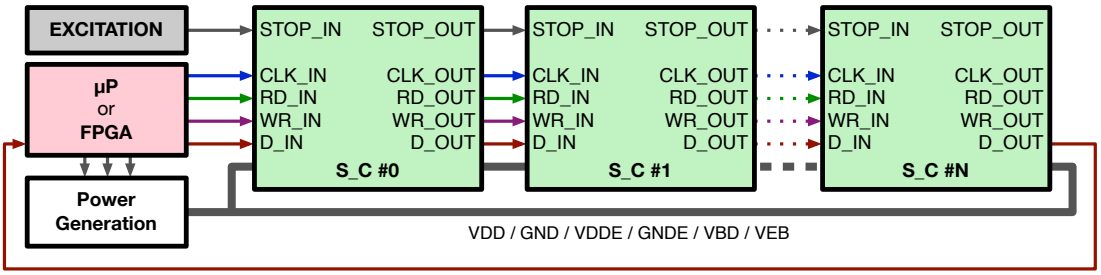


Figure 4.20: *Full daisy-chain network configuration.*

The primary concern with chaining devices in this way is the distribution of power supplies through the daisy chain, which will cause power *droop* towards the centre devices. As can be seen in Figure 4.1, the four core power supplies (From top to bottom: V_{BD}, V_{EB}, V_{DD} & *GND*) are all reinforced on non-light sensitive areas using the top metal layer. The width of these reinforced power straps is 50 µm, providing a typical resistance of $< 1\ \Omega$ per supply per chip (sheet resistance = 48 mΩ, length $\approx$ 1 mm). This is reduced even further with the inclusion of the power straps distributed within the pad ring. A worst-case current draw of 1 mA will cause a 1 mV drop per device, allowing up to 100 devices to be networked with only a 50 mV droop seen at the middle so long as power is provided at both ends of the chain, as

shown in Figure 4.20. Droop on the I/O power supplies (V_{DDE} and *GNDE*) will not have any effect on the detection or timing of photon events and can withstand dropping by as much as 500 mV to 2.8 V, so the pad ring power straps are sufficient in this instance.

A second concern with chaining devices, that is a particular issue for this application, is the buffering of the short excitation synchronisation pulse through all devices. The short pulse width, which varies from laser-to-laser could easily be lost to parasitics traversing the 1 mm width of a device. Therefore, the first device in a chain will be configured to route the pulse-lengthened *STOP* signal, as explained in Section 4.7, to its buffered *LASER_STOP_OUT* pad. Finally, a timing offset will be introduced by the delay of the *STOP* signal propagating through devices in the chain. By independently configuring the delay-lines in each device to compensate for the offset, its effect can be minimised.

4.7 Delay Line

4.7.1 Laser Synchronisation Delay

Typical TCSPC experimentation requires an external delay-line to allow the fluorescence decay to be positioned within the timing window of the TDC and also as a technique to improve jitter performance by delaying the *STOP* synchronisation signal until after the emission of the photon that its laser pulse created, as described in Section 2.2.2. A delay line is implemented on-chip to perform this functionality using a long buffer chain of 128 elements, as shown at the top of Figure 4.21, where each element is sized to provide $\approx$ 1 ns delay. The delay is selected using a hierarchical 128:1 multiplexer configured by a 7-bit register value.

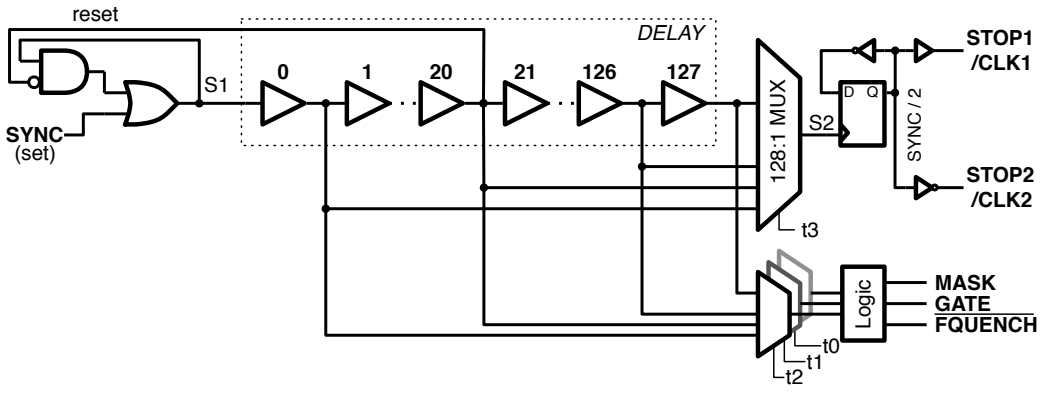


Figure 4.21: *Delay-line with pulse-lengthening SR-latch and multiplexers.*

Simulations showed the short pulse provided by the excitation source disappeared through the delay chain. One solution to this problem is to move the toggle flip-flop that divides the short synchronisation pulse (see Figures 4.11 and 4.14) to the front of the delay line, however transistor mismatch causes the divided clock signal to alter its duty cycle as it passes through the buffer chain. This in turn creates timing offsets between the two TDCs in each pair, which gets progressively worse as the delay is extended. To overcome these issues, an SR-latch is positioned at the input of the delay chain to extend the *SYNC* pulse duration, whilst the toggle flip-flop is positioned after the multiplexer. The SR-latch is *set* using the *SYNC* signal and *reset* using a fixed output from the delay-line at $\approx$ 20 ns, as shown in Figure 4.23.

4.7.2 SPAD Gating

It is of benefit to disable the SiPM array for a short period of time to make it insensitive during the optical excitation pulse, before quickly re-arming it to capture the fluorescence. Doing so would negate the requirement for an emission filter, which is particularly important to reduce the space between an unfocussed or uncollimated excitation source and the detector when the system is used without a microscope [34]. This functionality is implemented by creating a *MASK* signal to stop the output of the SiPM reaching the TDCs (see Figure 4.8). However, SPADs are still active during this time and have a chance of being *dead* when the SiPM is unmasked. Therefore, two further signals are included: *GATE* and $\overline{FQUENCH}$ (see Figure 4.3) disable then quickly re-arm the SPAD. An example of the timing of these signals is shown in Figure 4.22. Control of these signals is achieved by adding three additional multiplexers to the output of the delay line as shown at the bottom of Figure 4.21.

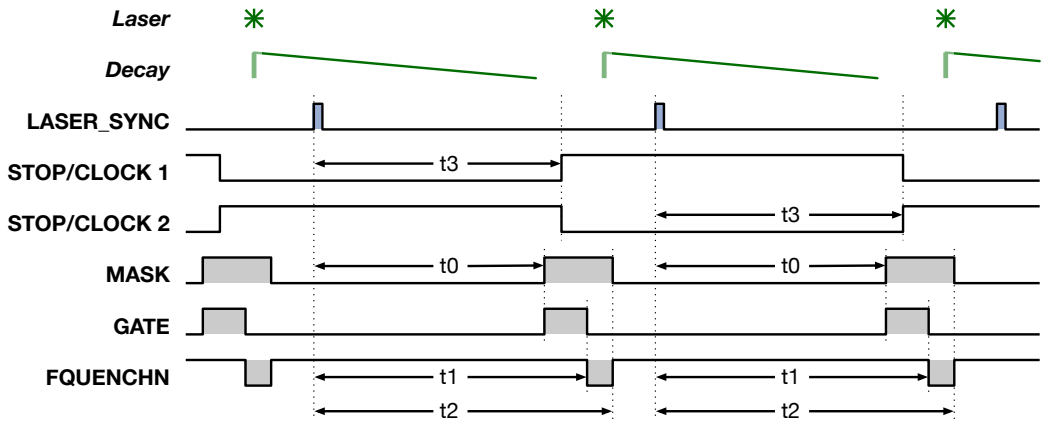


Figure 4.22: *Timing of delay generator signals.*

4.7.3 Verification

Simulations of the delay generator were performed with V_{EB} = 3.3 V, as the results in Figure 4.23 show. The *GATE* and $\overline{FQUENCH}$ signals are shown being level shifted to this value to be distributed using a clock tree into the SiPM array pixels. The delay register values were set to 64, 68, 70 and 66 for *t0*, *t1*, *t2* and *t3* respectively. Corresponding delay values of 55.7 ns, 59.4 ns, 60.9 ns and 57.1 ns are measured from the simulation, giving an approximate delay cell resolution of 0.87 ns, marginally bellow the specification of 1.0 ns.

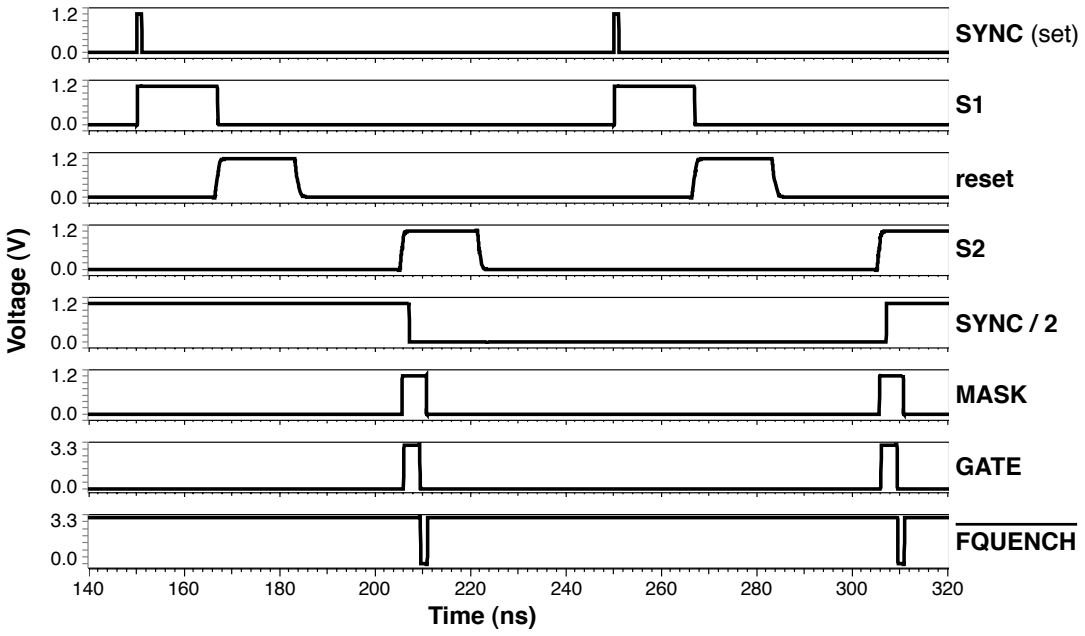


Figure 4.23: *Simulation of delay generation signals.*

4.8 Design for Test and Calibration

4.8.1 Standard TCSPC Operation

To allow the device to be configured to operate as a standard TCSPC sensor, a ninth TDC pair is added to the system architecture (as shown in Figure 4.2). This additional TDC pair can be configured and used in a number of ways, providing further functionality along with the ability to test, characterise and calibrate aspects of the system.

In the first instance, the output of the TDC pair is routed directly to 10 parallel output pads. The 10 output bits can be selected from any set of 10 ordered bits from the 16-bit TDC data bus using a multiplexer, with the finest resolution ($\approx$ 50 ps) providing over 50 ns total range and the largest resolution of $\approx$ 3.2 ns providing over 3.2 µs total range. These raw TDC codes can then be sampled off-chip at the excitation rate and so provides a system similar to current TCSPC modules, with the ability to time at most one event per excitation cycle, but with the added benefit of having no processing dead-time. These events can then be histogrammed in software or on FPGA to produce lifetime decays. As well as being used for experimentation, acting as a comparison to typical TCSPC acquisition set-ups, this mode is crucial for test, characterisation and calibration of the system.

Additionally, using the same 10 output pads, the TDCs can be configured to count the number of photon events within each excitation cycle (rather than a single photons time of arrival), as shown in Figure 4.10. By using a continuous wave (CW) laser rather than a pulsed laser and inputting a known time-base waveform into the *LASER_IN* port of the device, this mode can be used to record photon intensity traces, which are necessary for applications such as intensity scanned microscopy imaging and fluorescence correlation spectroscopy (FCS).

Bonding and connecting these 10 output pads for every device in a network of multiple devices is clearly not scalable. It was therefore also necessary to provide a method to read raw TDC codes using the serial interface I/O path (see Section 4.6). This was achieved by designing the serial interface to be configured to read out the 4 most recent TDC codes (4×16 = 64 bits) in place of the CMM data. However, this technique will only be able to produce TDC codes at an absolute maximum rate of 4 MHz, compared with the raw parallel data output that will run at the laser excitation frequency, which can be over an order of magnitude faster.

Finally, the ninth TDC pair can be used as the input to the CMM pre-calculation, in place of the multiple TDC architecture, by including a comparator, first stage summation and ripple counter as described in Section 4.5.2. This provides a means of direct comparison between one and eight timing channel CMM calculation. Additionally, a multiplexer can route the raw SiPM output into this ripple counter in place of the *VALID* signal, providing a large (28-bit) on-chip counter to perform long exposure intensity and DCR analysis.

4.8.2 TDC Calibration

Process, voltage and temperature (PVT) variations will create performance disparity across many aspects of the device, most importantly with the resolution, linearity and mismatch of the multiple TI-TDC architecture. The resolution in particular of the TDCs must be known at all times to allow for correct calculation and normalisation of both CMM and raw TCSPC data. Due to area constraints, on-chip PLL-style calibration, as is implemented on [12], is not an option. As such, an external off-chip method to calibrate for TDC resolution variation must be implemented. At the expense of an additional input pad (with corresponding output for networking, as described in Section 4.6.2) and one control register bit, the output of the SiPM can be multiplexed with a *TEST_START* signal which is designed to mimic a photon arrival at the input to the time-resolving circuitry. By using this signal together with a controllable *STOP* signal, it is possible to sweep their time difference and use the single TDC pair with raw outputs to build up a histogram. These histograms can then be used to graph the resolution and linearity of the TDC pair. The *TEST_START* signal is also routed to the front end of the token-passing event distribution, so that the multiple TDC architecture can be tested using known waveforms.

A second technique to measure the resolution and linearity of the TDC is to use a white-noise source (either uncorrelated light or SPAD DCR) with a known excitation period that is shorter than the full range of the TDC [137]. By capturing and histogramming raw TDC codes, the ideally flat white noise will fall off around the excitation period and we can use its known time to calculate the TDC resolution. The flatness of this histogram will provide information about the linearity of the TDC. In the event that the excitation source is producing the synchronisation signal, another input is required as a user controlled *TEST_STOP* source so that calibration can take place without the need to physically disconnect or connect wires. Therefore, the serial interface's *CLK* signal is routed to a multiplexer and can be used as the stop signal when the device is in such a situation.

4.9 Conclusions

This chapter has presented the design and implementation of the *SIPM_CMM* test chip to demonstrate the feasibility of high throughput fluorescence lifetime sensing. Combining a smart compressed output SiPM with an 8-channel TI-TDC pair architecture will allow the pile-up limit to be increased by over an order of magnitude over current single channel techniques. The sensor is believed to be the first implementation of an embedded fluorescence lifetime estimation pre-calculation, providing results in *real-time* to facilitate applications such as flow cytometry. The system miniaturisation is completed with the inclusion of an on-chip delay line, making this a truly integrated System on Chip (SoC). Thanks to the advanced 130 nm CMOS process and design for low I/O requirements, the device area is kept below 2 mm^2.

The 1024-element smart SiPM sensor provides a high throughput photon rate by compressing the output pulse-width of individual single photon sensitive detectors to 250 ps. This short pulse width enables a throughput in excess of the excitation frequency for fluorophores with lifetimes greater than 2.5 ns, a factor of two improvement over the expected performance presented in Section 3.8.3. Furthermore, individual detection elements in the SiPM can be enabled or disabled independently. As well as allowing individual SPADs to be characterised individually, this functionality can be used to switch off noisy detectors to improve SNR at the expense of reduced sensitivity. Additionally, an isolated SiPM power supply, V_{EB}, that is controlled independently of the core power supply voltage can be used as a sensitivity control. A comparatively high fill factor of 10 % has been achieved by moving circuitry to the periphery of the SiPM.

To negate processing dead-time, TDCs are time-interleaved in pairs, operating on alternate laser synchronisations. The multiple-channel timing per excitation is achieved by distributing SiPM output events to an array of eight TDC pairs using a token-passing circuit that has been designed to be immune to the metastability issues caused by the high speed asynchronous nature of photon events. However, this stability is achieved at the cost of increased CMM calculation error caused by the use of a *resetting* router which accentuates TDC mismatch, as described in Section 3.10. This is expected to present the primary source of calculation error in the manufactured system. Extending the range of the TDC to 16 bits, or > 3 µs, significantly increases the range of fluorophores that can be measured using the system, enabling Oxygen sensing using RTDP [16, 38].

Not only does embedding the CMM pre-calculation provide fluorescence lifetime estimations in *real-time* but it negates the requirement for a high-speed, highly parallel output bus to transfer the data from all 16 TDCs off-chip, keeping device power consumption and area low. A read-out dead-time is introduced to allow the pipelined adder tree and accumulator to settle, however steps have been taken to minimise this and the large data widths (28 and 36-bits) for the CMM pre-calculation allow long exposures to minimise this further.

A bi-directional shift-register serial-interface has been incorporated to keep the I/O requirements low for reading data off-chip and writing data to on-chip registers. The custom interface also facilitates the ability to network multiple devices together to reduce system overheads in a multi-detector set-up. A delay line is embedded on chip, completing the miniaturisation of the detection path of a TCSPC set-up. Not only does the delay line allow the histogram to be positioned within the TDC's timing window, but can also create user controlled signals to gate the SiPM array during excitation to allow the emission filter to be removed from the optical path.

The device has been designed with a number of techniques to test and characterise the SiPM, timing and CMM performance that will be used extensively in Chapter 5. Furthermore, these techniques can be used to perform DCR and timing calibration that are necessary to correct for system non-idealities when using the device in a fluorescence lifetime experiment. The MATLAB model introduced in Chapter 3 will be adapted and re-run with the final design parameters to compare with practical fluorescence lifetime experimental results. Finally, Chapter 6 will introduce implementation refinements and changes that would improve the performance of the sensor.

5
SENSOR TEST AND CHARACTERISATION

5.1 Introduction

This chapter presents the characterisation of the *SiPM_CMM* sensor reported in Chapter 4. The device is tested and characterised electrically and optically as well as being used in practical fluorescence lifetime microscopy experimentation. The chapter begins by describing the evaluation platform, which includes a printed circuit board (PCB), field programmable gate array (FPGA) firmware and software to facilitate the entirety of the testing, characterisation and experimentation procedures.

Following the description of the evaluation platform, the characterisation of the individual components of the sensor – SiPM, timing and delay line – are presented. Next, the system is characterised as a whole, investigating the instrument response and device power consumption. The sensor is then characterised for raw TCSPC and high-throughput CMM data capture with bulk sample fluorescence lifetime experiments. Finally, the device is used to perform scanned FLIM and simulated flow based experiments. The chapter finishes with a critical discussion of all the presented results.

5.2 Test Platform

5.2.1 System Overview

The evaluation platform is built around an *Opal Kelly*[1] hardware module that incorporates an FPGA, USB communication and other peripherals. To facilitate fast hardware prototyping, pre-compiled FPGA modules and software APIs are available to perform the low-level USB communication protocols. A block diagram of the full system is shown in Figure 5.1, with the *Opal Kelly* modules highlighted in grey. In addition to the *SiPM_CMM* sensor, the primary custom components of the system are the hardware in the form of a PCB (blue), FPGA firmware (red) and software (green). Each of these primary components will be explained in detail in the following sections.

[1] Further information, data-sheets and user-manuals available from http://www.opalkelly.com.

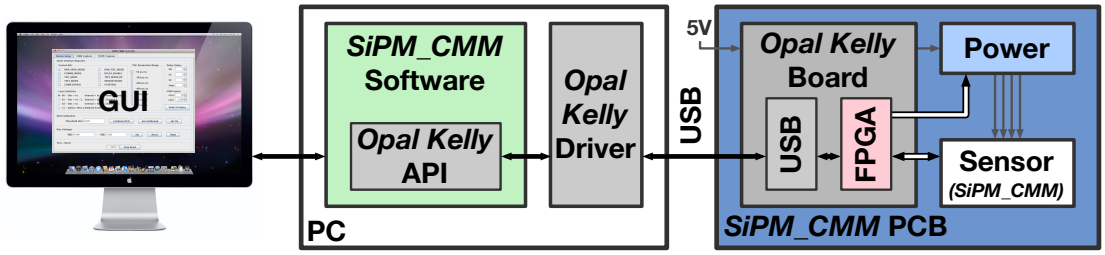


Figure 5.1: *Block diagram of* SiPM_CMM *evaluation platform.*

5.2.2 Hardware

As shown in Figure 5.2, the hardware platform for the *SiPM_CMM* sensor consists of three primary components: an *Opal Kelly* plug-in FPGA/USB module, a custom PCB and a packaged *SiPM_CMM* die. The main component on the plug-in module is a *Xilinx Spartan 3* FPGA (XC3S1500), which provides sufficient resource for the compact firmware implementation. The device is packaged in a 48-pin ceramic lead-less chip carrier (CLCC) with transparent glass lid. Details of the four-layer PCB layout and schematics can be found in Appendix A.9.

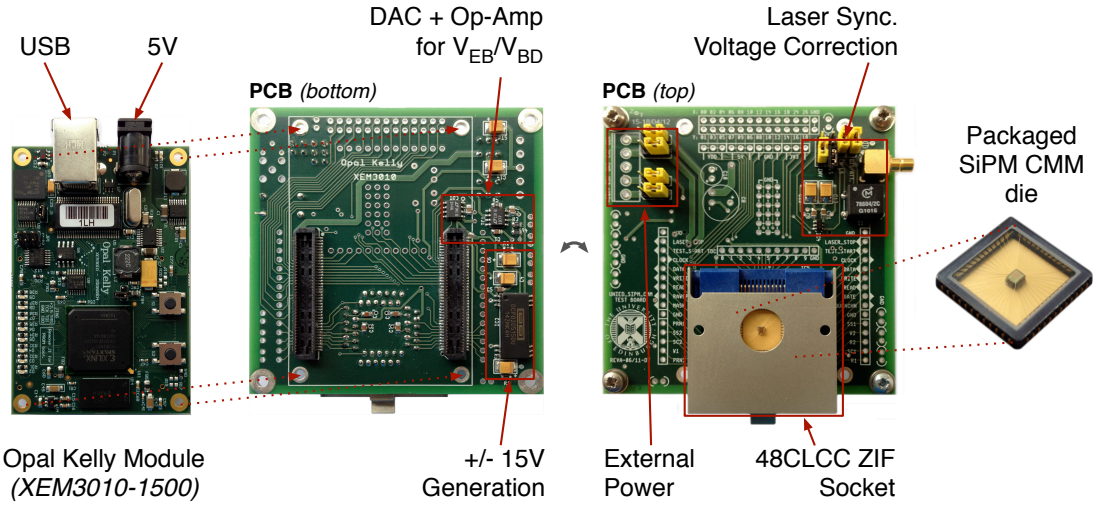


Figure 5.2: *Details of test platform hardware, showing* Opal Kelly *plug-in FPGA/USB board (left), custom PCB showing bottom and top (centre) and packaged die (right).*

As well as interfacing to both the plug-in FPGA module and the packaged sensor, the PCB contains two additional circuit blocks. The first of these is a set of components designed to generate user-controllable bias voltages for V_{BD} and V_{EB}. A dual output DAC (LT1446 - introduced in Section 4.6.1) is connected to the FPGA to allow two voltages to be created

between 0 V and 5 V. An operational amplifier (LT1077) configured with a gain of -4 is then used to allow one of the DAC outputs to produce the V_{BD} bias voltage. A DC/DC converter (DCP020515D) is used to produce the +/-15 V power supplies for the op-amp from the single 5 V power supplied to the plug-in module. The plug-in module also contains low-dropout regulators that provide 1.2 V and 3.3 V, amongst others, that are used for V_{DD} and V_{DDE}, respectively. Additionally, the PCB can be configured to allow all four power supplies to be independently sourced externally. Furthermore, the selection jumpers to achieve this have the dual functionality of allowing ammeters to be placed in each supply path in order to monitor the system current draw to calculate power consumption.

The second functionality included on the PCB is a circuit designed to accept a laser synchronisation pulse conforming to the nuclear instrument module (NIM) standard (-16 mA into 50 Ω = -800 mV), from which the majority of pulsed lasers are designed. This signal must be converted to a CMOS compatible signal (+3.3 V) to allow it to operate with both the *SiPM_CMM* device and the FPGA. The conversion is achieved using a pulse transformer (78604/2C) to invert the signal and a level-shifter (SN74AVC1T45). Additionally, some synchronisation signals conform to a TTL standard (5 V), so an optional potentiometer is also included that can be configured independently from the NIM circuit. A number of selection jumpers are available to achieve these different operation modes.

5.2.3 FPGA Firmware

In order to drive the various control signals required by the sensor, to receive and process the output data from it, and to package the processed data for transmission to a PC for further analysis, an FPGA or micro-controller containing custom firmware and communication functionality is required. As introduced previously, an off-the-shelf plug-in FPGA/USB module is used to realise these implementation requirements. The plug-in board provides pre-compiled firmware to implement the low-level USB communications and handle the FPGA configuration process, significantly shortening the firmware development time. The pre-compiled firmware modules provide a simple register read/write interface, triggering functionality and high-speed bulk data transfers. A simplified module-level architecture of the firmware required for the *SiPM_CMM* device is shown in Figure 5.3, the individual blocks of which will be expanded on in the remainder of this section. Descriptions of the signalling protocols required for serial interface communications are given in Section 4.6.1 and Appendix A.7.

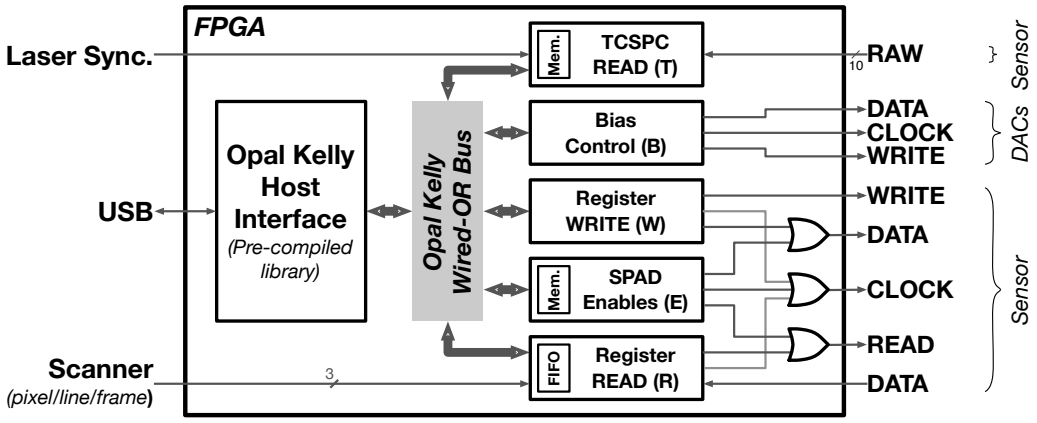


Figure 5.3: *Custom FPGA firmware architecture.*

Serial Interface *WRITE* functions (**W**) are implemented using a 64-bit on-FPGA register, which is accessible through the *Opal Kelly* interface's register write functionality, before using a finite state machine (FSM) to control the *WRITE*, *DATA* and *CLOCK* output signals. Configuring the SPAD enable signals (**E**) is performed in a similar way, with the 64-bit on-FPGA register being replaced with a 1024-bit memory, also accessible through the *Opal Kelly* interface's register write functionality and controlling the *READ* output in place of *WRITE*.

A first-in first-out (FIFO) buffer is used to allow many sequential serial interface read functions (**R**) to be performed in quick succession, without the available USB bandwidth being compromised. The module controls the *READ* and *CLOCK* outputs, whilst sampling data arriving at the *DATA* input from the sensor. As the FIFO fills up, it is bulk transferred through the *Opal Kelly* interface at speeds of up to 20 MB/s. When used with a scanning system, the transmission of each *packet* of data is synchronous with the *PIXEL*, *LINE* and *FRAME* clock inputs. Otherwise the exposure time and number of sequential reads are controlled using FPGA registers. Raw data TCSPC reads (**T**) are performed by sampling the incoming 10-bit *RAW* data bus from the sensor on every excitation synchronisation pulse. The incoming TCSPC codes are used to address and increment a counter in a 1024×16-bit memory which stores the histogram. At the end of an exposure, the length of which is set-up using a register write, the contents of the TCSPC histogram memory are bulk transferred through the *Opal Kelly* interface.

Bias levels (**B**) are accessible through the *Opal Kelly* interface and an FSM within the *Bias Control* block sets up the DACs accordingly. Some functionality, such as firmware and chip resets as well as some configuration/control registers, are not shown in Figure 5.3.

5.2.4 Software

A software application is required to interact with the FPGA to configure settings and acquire data from the sensor for further processing and analysis. It was decided to build a graphical user interface (GUI) to allow the software to be easily used and to provide instant visual feedback of user interactions and captured results. The JAVA programming language was chosen for the project for a number of reasons: a Java API is available from *Opal Kelly*, it is cross-platform (works with Windows, Linux and OS X) and previous experience with hand-crafting GUIs using the Java *Swing* toolkit meant that development time was saved building the application.

To keep the *SiPM_CMM* application software uncomplicated to develop, understand and maintain, it is built with a variation of the established Model View Controller (MVC) architecture [138], which partitions the software into three distinct components, most importantly separating the functionality (*Model*) from the GUI (*View*). The *Model* is concerned only with the the operational aspects of the system, for example how to write to and read from the FPGA. Conversely, the *View* is only used to present an interface to the user and contains no inherent knowledge of the underlying system. The *Controller* then deals with passing requests from the *View* to the *Model* and returning results from the *Model* back to the *View* for visualisation, if required. A Java class is created to implement each of the three functions in the MVC architecture. A visual representation of the software architecture is shown in Figure 5.4, highlighting the interaction between classes and providing example function calls in the *Model*.

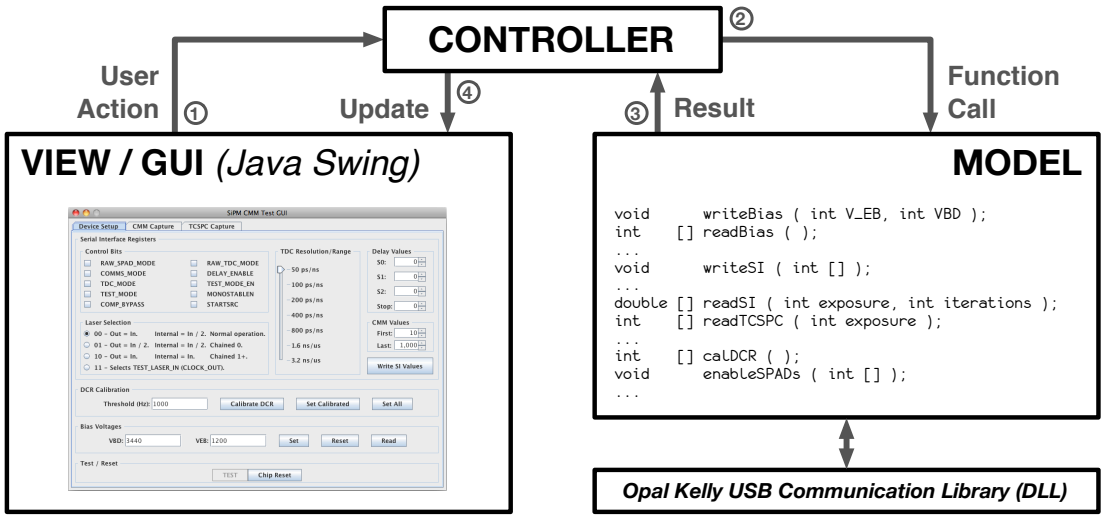


Figure 5.4: *Visual representation of* SiPM_CMM *software architecture.*

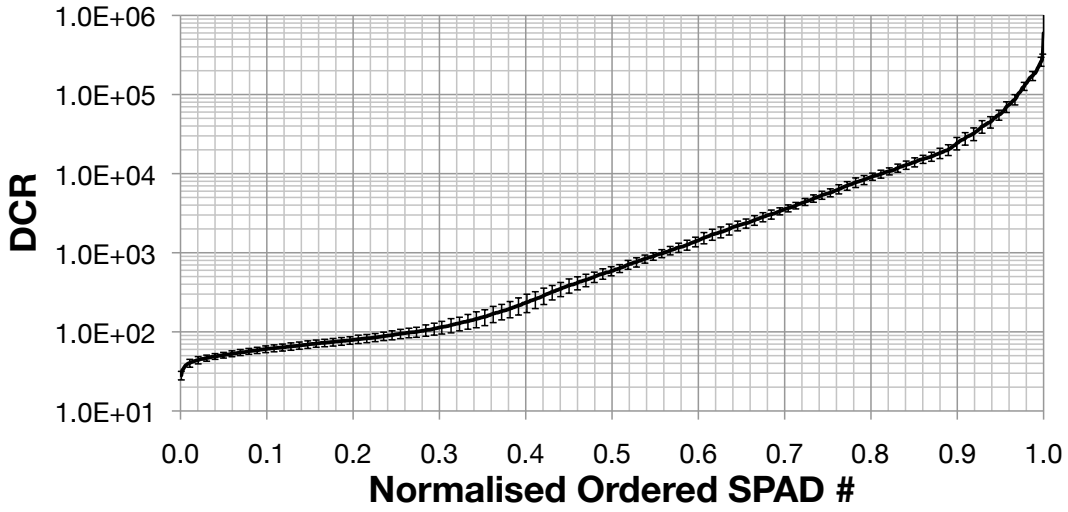
5.3 SiPM Characterisation

5.3.1 Overview

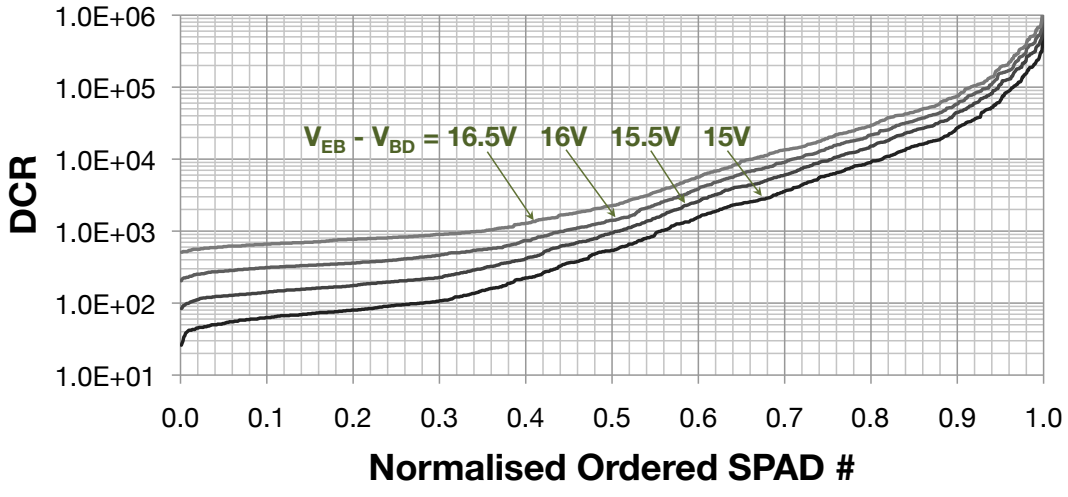
Bring-up and testing of the *SiPM_CMM* device begins with a characterisation of the single photon sensitive SiPM architecture (see Section 4.3). This characterisation starts with measurements of the SPAD DCR distribution and the effect that bias conditions have on it. These measurements also provide useful information on the spatial distribution of DCR within the SiPM. Next, the power consumption is reported for a varying number of enabled SPADs and increasing incident light intensity. Finally, an investigation of the available photon throughput is performed with the monostable disabled and enabled, the light levels and bias conditions varied, and the number of enabled SPADs increased from 1 to 1024. The DCR and throughput measurements are captured using automated software routines that use the 28-bit on-chip counter described in Section 4.8.1. A selection of SPADs are also output directly from the chip and monitored on an oscilloscope (*LeCroy WaveRunner 64MXi-A 600MHz 10 GS/s*) to validate the measurements. However, raw SPAD output can only be monitored when the monostable is disabled due to pad bandwidth limitations, so the on-chip counter is necessary when it is enabled.

5.3.2 Dark Count Rate (DCR)

The DCR distribution of the SPADs within the SiPM is measured in a controlled dark-room environment by enabling each element individually and capturing a long (ten second) exposure of the pulses produced. The results from this measurement are ordered from lowest to highest DCR, resulting in the graphs shown in Figure 5.5. Figure 5.5a shows the mean distribution and standard deviation of six measured devices for a fixed total bias (V_{OP}) of 15 V, and Figure 5.5a shows the effect of increasing the bias voltage on the distribution for one device. The distribution is noticeably worse than [12], as shown in Figure A.1, with only $\approx$ 30 % of SPADs showing a sub 100 Hz DCR. This can be partially attributed to a move from 200 mm to 300 mm wafers for fabrication and the increased active area of the SPAD, which has a negative impact on DCR by increasing the likelihood of a given device containing a silicon defect. However, as proposed in Chapter 3, only 16 detectors are required to achieve the specified throughput improvements, so finding a suitable compact region of detectors with low DCR is a viable option. As expected, increasing the bias conditions also causes the DCR to increase significantly, with the advantage of an increase in photon detection probability [59].



(a)



(b)

Figure 5.5: *Ordered DCR distributions showing* **(a)** *standard deviation at fixed bias conditions and* **(b)** *effect of varying bias conditions.*

To study the possibility of using a suitable region of 16 adjacent low DCR SPADs for experimentation, the results are plotted according to their location within the SiPM for one device. This spatial analysis is shown in Figure 5.6, where the DCR values are plotted on a logarithmic false-colour scale. There does not appear to be any spatial correlation of DCR, with the noisiest detectors being randomly distributed across the array. These noisiest detectors should be avoided for any practical experimentation. The three regions outlined in the figure highlight possible candidates of 16 adjacent SPADs for experimentation, providing relatively low total DCRs of 1.7 kHz (**A**), 1.2 kHz (**B**) and 1.4 kHz (**C**).

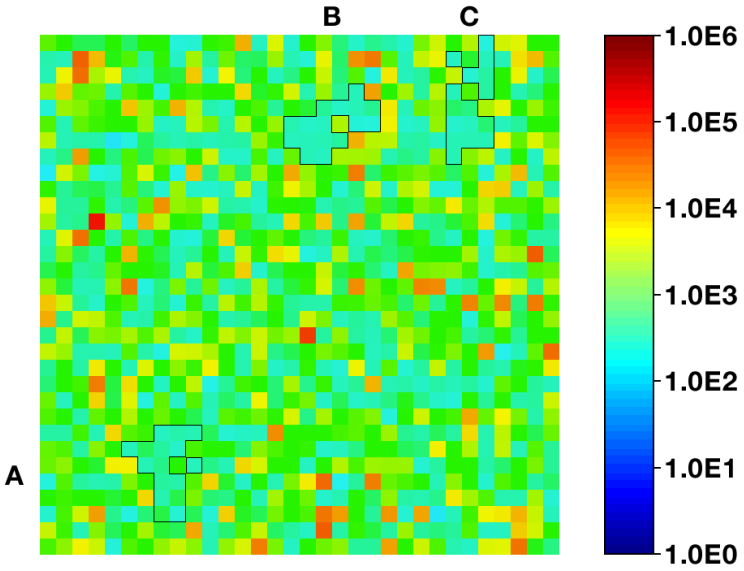


Figure 5.6: *Positional DCR for single device on log scale.*

5.3.3 Power Consumption

The SiPM power consumption is shown in Figure 5.7 as a function of incident light intensity (optical power) and for different numbers of enabled detectors. It is measured with a fixed bias voltage (V_{OP}) of 15 V. The SiPM bias voltages also draw power for the buffering and level shifting of detector output signals between voltage domains, so this is included in the measurements presented. The power consumption plateaus at 500 µW for the nominal 16 detector configuration, which is only 5 % of the specified 10 mW for the entire device (see Sections 1.3 and 4.1.2), leaving sufficient head-room for the multiple channel time-interleaved time conversion and embedded processing. The effect of enabling additional detectors is significant – enabling all 1024 causes a peak power consumption of over 10 mW.

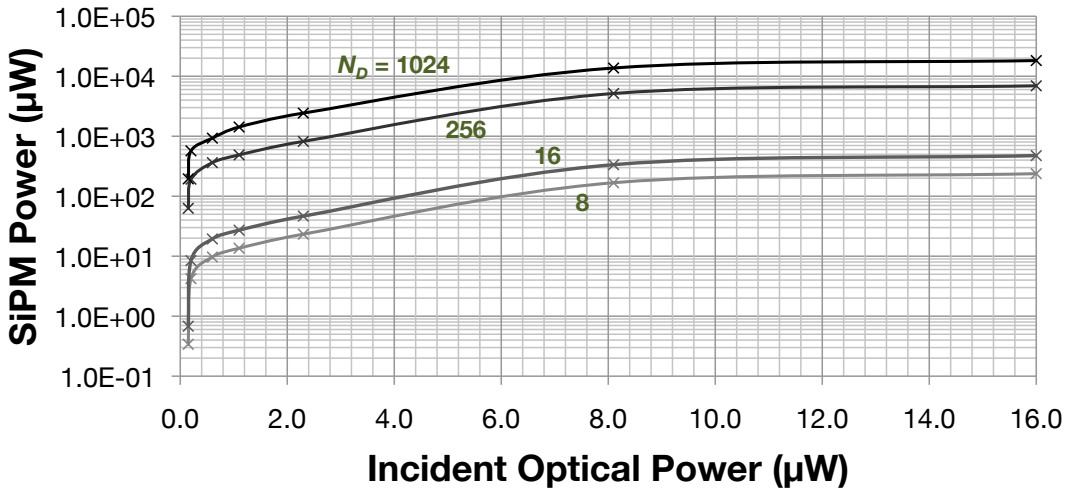


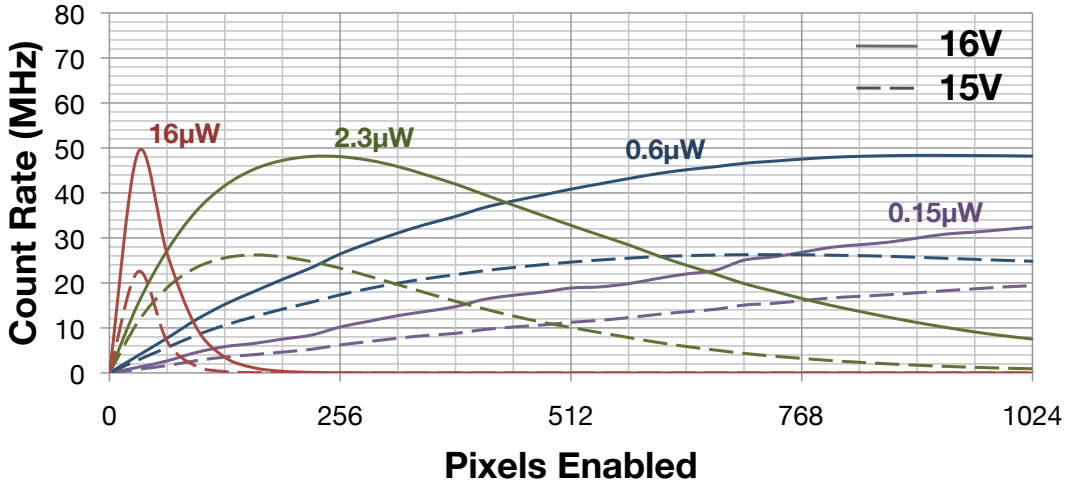
Figure 5.7: *SiPM Power Consumption.*

5.3.4 Throughput

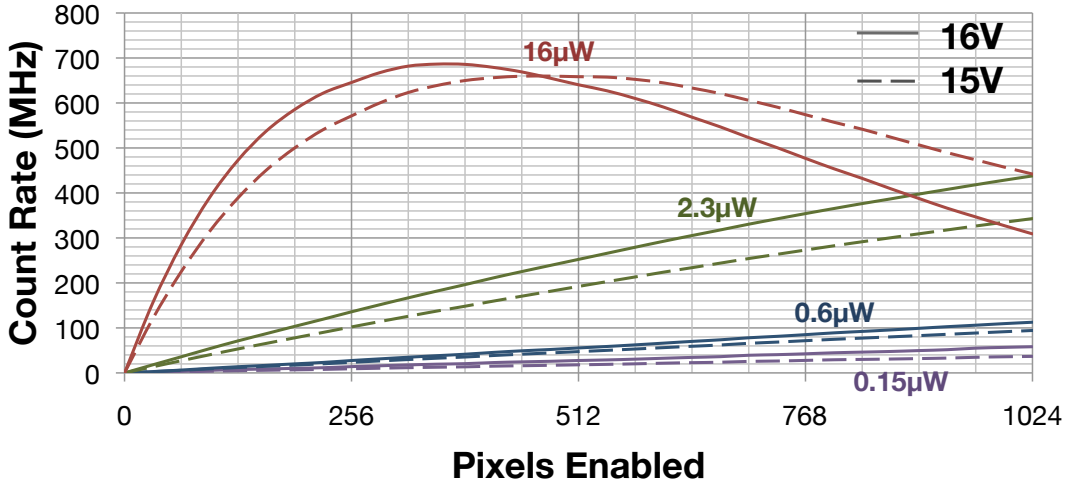
The photon throughput of the SiPM is of critical importance to the successful operation of the device architecture as it will define its limitations in terms of both detector dead-time and channel pulse overlap pile-up, as introduced in Section 3.2.2. The results presented in this section will aim to confirm the performance gains available by incorporating the pulse shortening monostable circuit at the output of each individual detection element. This is achieved by measuring the count rate produced at the SiPM output for an increasing number of enabled detectors between 1 and 1024. Increasing the number of detectors in this way provides a gain control that will affect the maximum count rate achievable. The tests are run using various operating conditions such as light level, bias voltage and with the monostable enabled and disabled.

The test is initially performed with the monostable *disabled* to highlight the limitations of SPAD dead-time on the available photon throughput. The graph in Figure 5.8a shows the results from this test for fixed light levels of 16 µW (red), 2.3 µW (green), 0.6 µW (blue) and 0.15 µW (purple) and for bias voltages of 16 V (solid) and 15 V (dashed). The primary effect of decreasing the bias voltage with the monostable disabled is to increase the SPAD dead-time, which in turn causes a significant drop in the maximum achievable photon rate from $\approx$ 50 MHz to $\approx$ 25 MHz due to increased pulse-overlap in the channel.

The results of performing the same experiment, but with the monostable *enabled*, are shown in Figure 5.8b. The total throughput is now as high as 700 MHz, over an order of magnitude higher than with the monostable disabled (note the scales of the y-axes in Figure 5.8). The primary effect of decreasing the bias voltage with the monostable enabled is to decrease the PDE of the SPAD detectors, which reduces the slope of the count rate gain. In this case, although the SPAD dead-time is still increased, it has a negligible effect due to the much shorter monostable pulse-width.



(a)



(b)

Figure 5.8: *SiPM count-rate response with varying light levels and increasing number of enabled detectors for monostable circuit disabled* **(a)** *and enabled* **(b)**.

Due to the use of a passive quenching element in the SiPM implementation, the results can be fitted to the paralysable detector model presented in [99]: $m = n \cdot e^{-n \cdot t_D}$, where m is the measured count rate, n is the true count rate and t_D is the detector dead-time, or channel pulse-width in this case. The true count rate (n) in both cases is calculated as $\approx$ 250 kHz/µW and $\approx$ 185 kHz/µW per detector for the 16 V and 15 V bias conditions, respectively. For the case where the monostable is disabled, the detector dead-times (t_D) are calculated as $\approx$ 7.7 ns and 11.1 ns for the 16 V and 15 V bias conditions, respectively. These calculated dead-times are verified by measuring the pulse width distribution of a raw SPAD output on the oscilloscope. With the monostable enabled, the SiPM output pulse-width is calculated to be $\approx$ 540 ps, independent of the bias voltage. This is over twice as long as was designed and verified from extracted layout in Section 4.3.3 and is attributed to extraction discrepancies.

Using the equation presented above, the graph in Figure 5.9 shows the efficiency of the SiPM with the nominal 16 detectors enabled for increasing optical power from 0.1 µW to 10 mW and with the monostable enabled and disabled. This figure describes the divergence from the ideal linear response caused by pile-up in the SiPM channel. The ideal count rate (n) is directly proportional to the optical power and is 3.84 MHz/µW and 2.96 MHz/µW for the 16 V and 15 V biases, respectively. The effect of reducing the bias when the monostable is disabled is negligible as the reduced detection efficiency and increased dead-time negate each other. With the monostable disabled, a 3 % loss of efficiency is recorded at just 1 µW (n = 3.84, 2.96 MHz), whilst enabling the monostable provides over an order of magnitude improvement (on the same order as the reduction in pulse-width) with only 2 % loss at 10 µW (n = 38.4, 29.6 MHz). The 16 detector SiPM is completely paralysed by 200 µW (n = 768, 592 MHz) and 3 mW (n = 11.52, 8.88 GHz) with the monostable disabled and enabled, respectively.

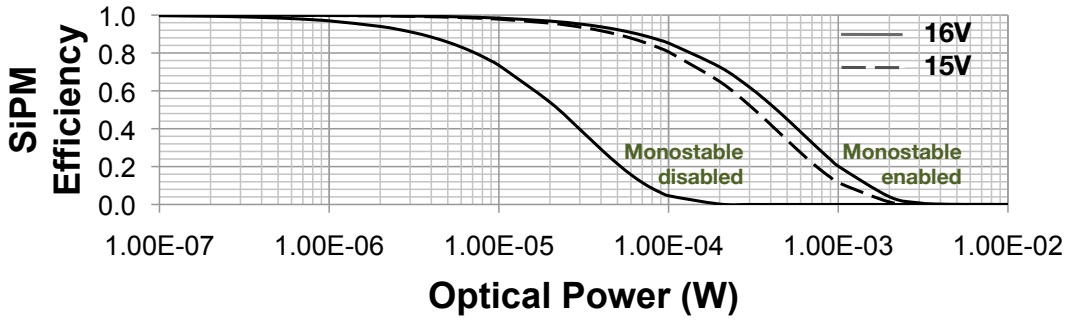


Figure 5.9: *Efficiency of SiPM with 16 detectors for increasing light level.*

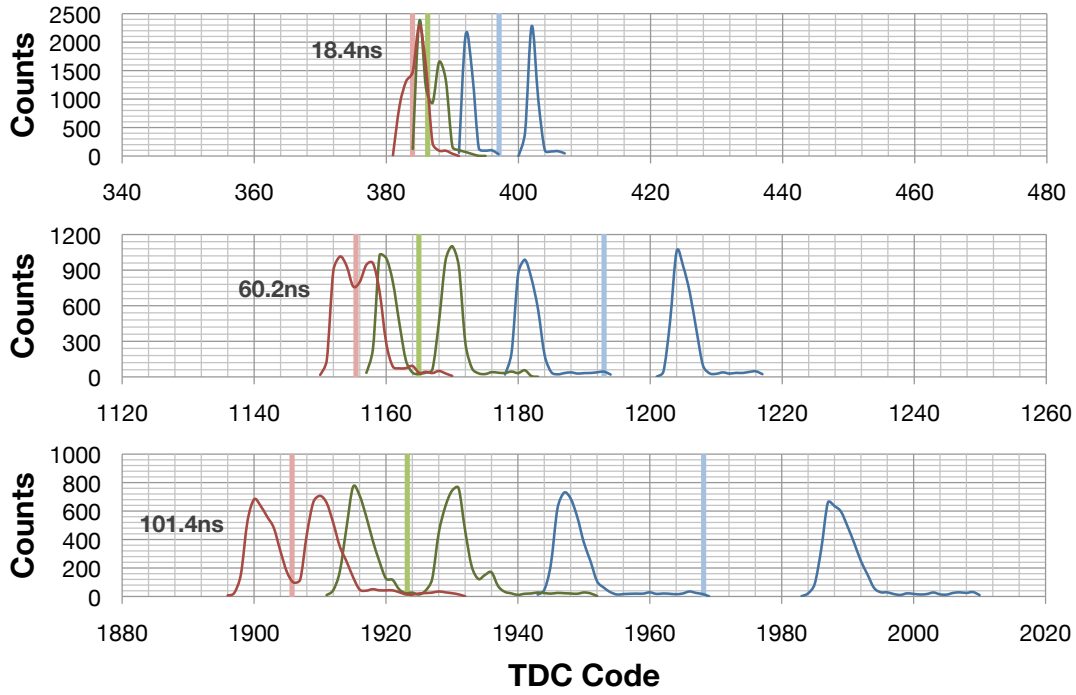
5.4 Electrical TDC Characterisation

5.4.1 Overview

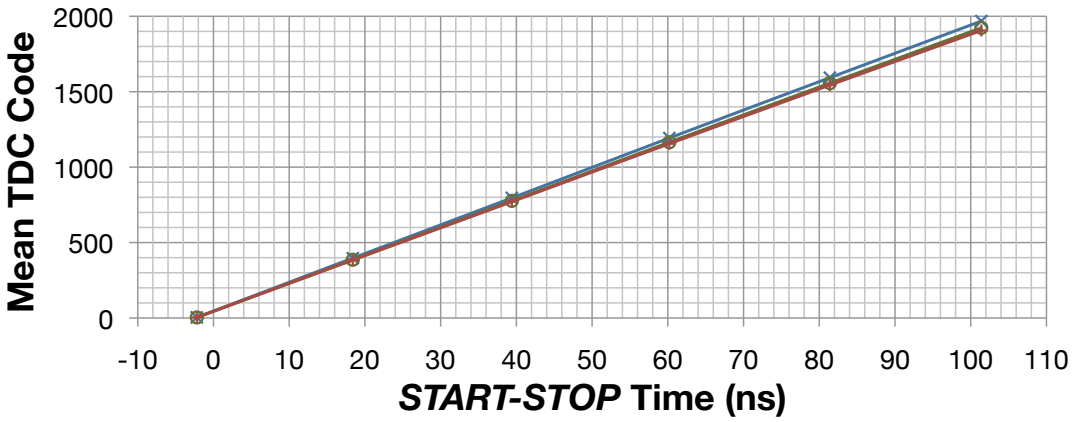
This section describes the characterisation and evaluation of the TDC timing performance using the test-mode and calibration techniques introduced in Section 4.8. In particular, the device is configured to use the test TI-TDC pair with known and controllable electrical *TEST_START* and *STOP* signals. All of the results presented within this section are captured at room temperature with the SiPM disabled and the *START* and *STOP* signals generated by the FPGA, with their timing verified using the oscilloscope. To begin with, the resulting TDC histograms of the electrical response to known fixed *START*-*STOP* durations are presented for three representative devices with varying TDC mismatch. The average TDC codes from these measurements are then plotted against the known *START*-*STOP* durations. Next, the full width at half maximum (FWHM) of the timing response is plotted as a function of *START*-*STOP* time for a single TDC and for the TI-TDC pairs, which includes mismatch. Finally, the effect of voltage variation on the TDC resolution is investigated, highlighting the need for calibration during experimentation.

5.4.2 Timing Response

The graphs in Figure 5.10a show the timing response of the test TI-TDC pairs of three devices for three fixed *START*-*STOP* times of 18.4 ns, (top), 60.2 ns (middle) and 101.4 ns (bottom). The measurements are captured with a *START*-*STOP* rate of 6.25 MHz and a fixed exposure of 1.28 ms, providing 8,000 samples per device (4,000 samples per TDC). The mismatch between TDCs is clearly apparent, particularly for the third device (blue). The graphs are shown on the same x-axis scale to highlight the widening of the individual TDC response and the increasing gain mismatch error between pairs. The average TDC code for each device is highlighted by the vertical lines in the figure, and these calculated averages are plotted against the known time difference to produce the graph shown in Figure 5.10b. The resulting gradients of fitting linear equations to these results describes average TDC resolutions of 54.3 ps, 53.8 ps and 52.7 ps for the three devices. Isolating the individual histograms of the device with worst case mismatch (blue) and calculating their mean values as a function of *START*-*STOP* time provides individual TDC resolutions of 52.2 ps and 53.2 ps, which match well with the expected standard deviation of $\pm$0.45 ps used for mismatch modelling in Section 3.10.



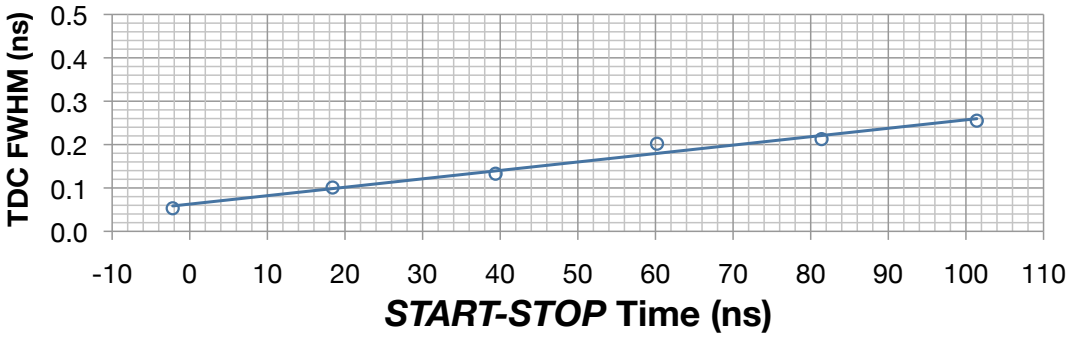
(a)



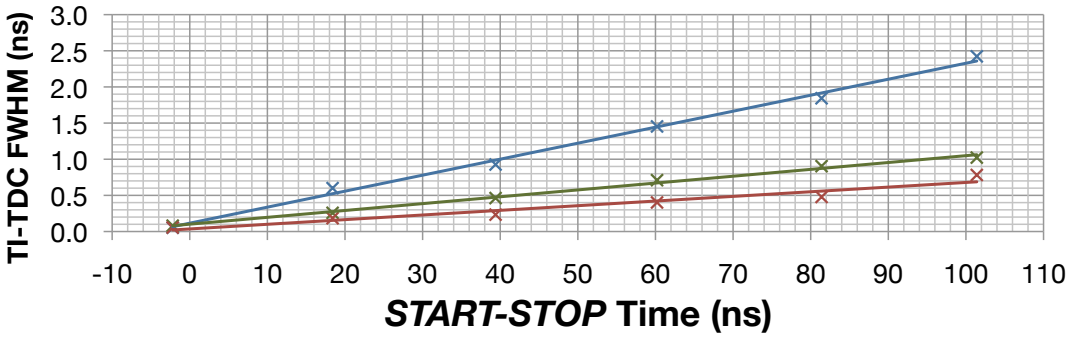
(b)

Figure 5.10: (a) *Histograms showing TDC response using electrical stimulation for three fixed time delays of 18.4 ns (top), 60.2 ns (middle) and 101.4 ns (bottom) with the average code highlighted in each case, and* **(b)** *the average code plotted as a function of the time difference between START and STOP. All results are plotted for three devices with varying mismatch (red, green and blue).*

As shown in Figure 5.10a, the FWHM of the individual TDCs gets larger as the input *START-STOP* time is increased. By isolating the histograms of the device with the worst-case mismatch, the FWHM of these individual TDCs can be plotted against the input *START-STOP* time, as shown in Figure 5.11a. Fitting this graph to a linear equation gives the following expression: $FWHM \approx 0.002 \cdot t + R_T$, where t is the input time and R_T is the resolution of the TDC. Therefore an input delay of 75 ns is required before the TDC FWHM begins to dominate the SPAD jitter of $\approx$ 200 ps. However, plotting the FWHM of the combined mismatch affected TI-TDC pair produces much higher values that are device dependent, as shown in Figure 5.11b. In this case at the same 75 ns input delay, the FWHM for the three measured devices are 0.5 ns (red), 0.7 ns (green) and 1.7 ns (blue). This performance is as expected and the modelling in Section 3.10 describes its effect on fluorescence lifetime calculation by CMM. It is not possible to capture raw histograms for the eight channel TI-TDC architecture, so the modelling is important to fully understand the CMM results presented later in this chapter.



(a)

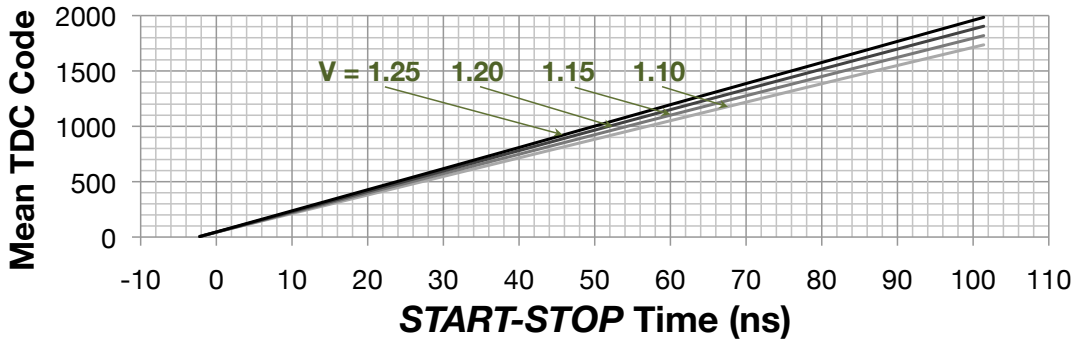


(b)

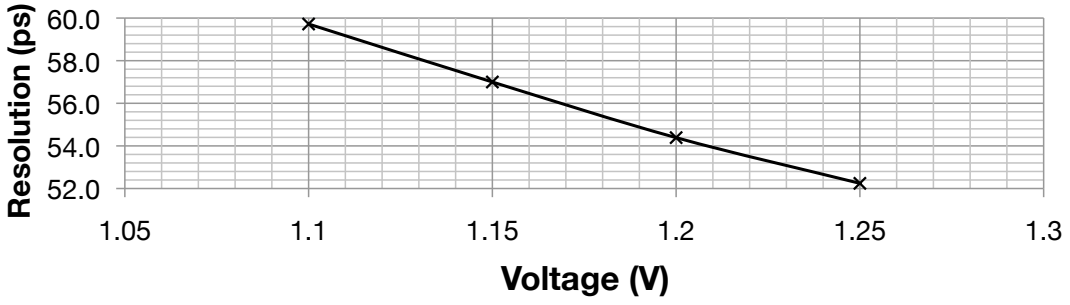
Figure 5.11: *FWHM of three devices as a function of mean code for* **(a)** *single test TDCs and* **(b)** *test TI-TDC pair.*

5.4.3 Supply Voltage Variation

Operating conditions such as process, voltage and temperature (PVT) will have a large impact on the TDC resolution. The effect of process variation in the form of mismatch has been presented in the previous sub-section (5.4.2). To investigate the effect of supply voltage variation, the average TDC code from one device is calculated for increasing *START-STOP* time and for varying supply voltage levels. The results from this are shown in Figure 5.12a for voltages of 1.10 V, 1.15 V, 1.20 V and 1.25 V (the device fails to operate outside of this window). The resolutions that these voltages produce are plotted in Figure 5.12b, showing a linear relationship between 60 ps and 52 ps for increasing voltages. The device is typically operated at a nominal 1.20 V, providing a TDC resolution of 54.4 ps at room temperature. Temperature variability is also expected to affect the resolution in a similar way, but is not measured due to the unavailability of a temperature controlled oven. All of these results highlight the necessity to calibrate the device *before* and also *during* long experiments, if possible.



(a)

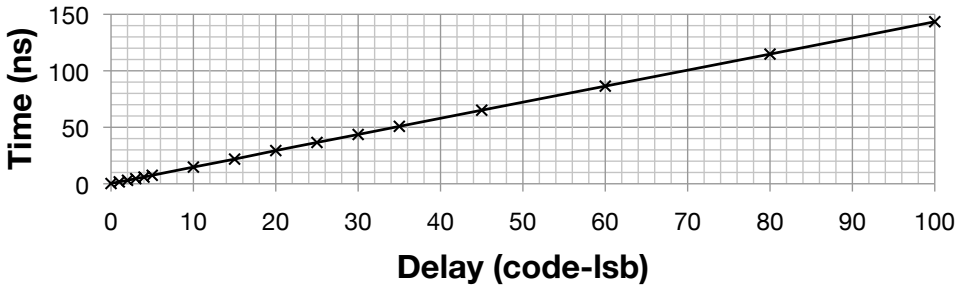


(b)

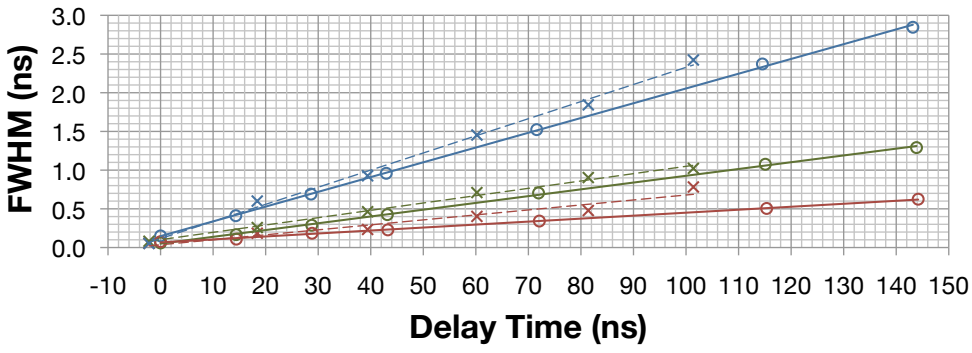
Figure 5.12: **(a)** *The average TDC code plotted as a function of the time difference between start and stop and* **(b)** *resulting TDC resolution, for varying core supply voltages (V_{DD}).*

5.5 Delay Line Characterisation

The final block within the device to be characterised is the embedded delay line. A similar investigation to the one described in Section 5.4.2 is performed, but using the delay line to provide the *START*-*STOP* time difference in place of the FPGA. The results are presented in Figure 5.13a, showing the time produced by the test TI-TDC with increasing delay code from 0 to 100. Fitting a linear equation to these results gives an LSB value of $\approx$ 1.43 ns and a full scale range of 128·1.43 ns $\approx$ 183 ns. This is 43 % larger than specification and 65 % larger than the simulated value from Section 4.7, following the same trend as the increase in the SiPM channel pulse-width presented earlier in this chapter. The results for all three devices are measured with a 1.2 V supply and at room temperature, providing an LSB value within 1 % error of each other. Measuring the FWHM of the individual and TI-TDCs using the delay line to set the input time difference shows a noticeable improvement over the FPGA approach in previous sub-sections. This is highlighted in Figure 5.13b and is made possible by a reduction in timing uncertainty (jitter) between *START* and *STOP* that is worse in the case of the FPGA approach.



(a)



(b)

Figure 5.13: **(a)** *The average time produced by the TI-TDC pairs plotted as a function of the embedded delay code and* **(b)** *the FWHM of the three devices as a function of the input time difference using the embedded delay (solid).*

5.6 System Characterisation (Instrument Response)

5.6.1 Overview

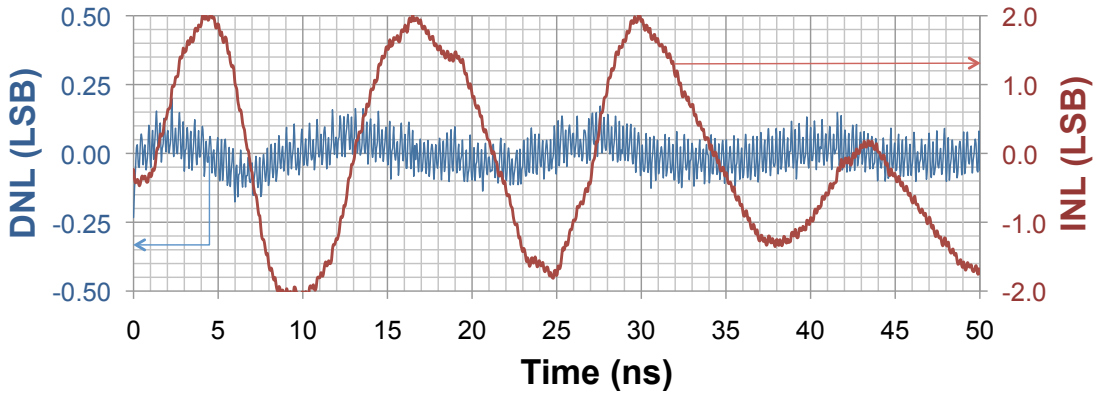
This section describes the characterisation of the system's instrument response using the SiPM and test TI-TDC with different device configurations. The investigation begins with a code-density test by measuring the response to an optical white-noise input. Next, the instrument (impulse) response function (IRF) is investigated, highlighting a SPAD position dependent timing variation that is studied in further detail. The IRF results are captured using unfocused optical excitation in the form of a pulsed diode laser. Finally, the power consumption of the device using two and eight channel TI-TDCs is presented.

5.6.2 Code Density

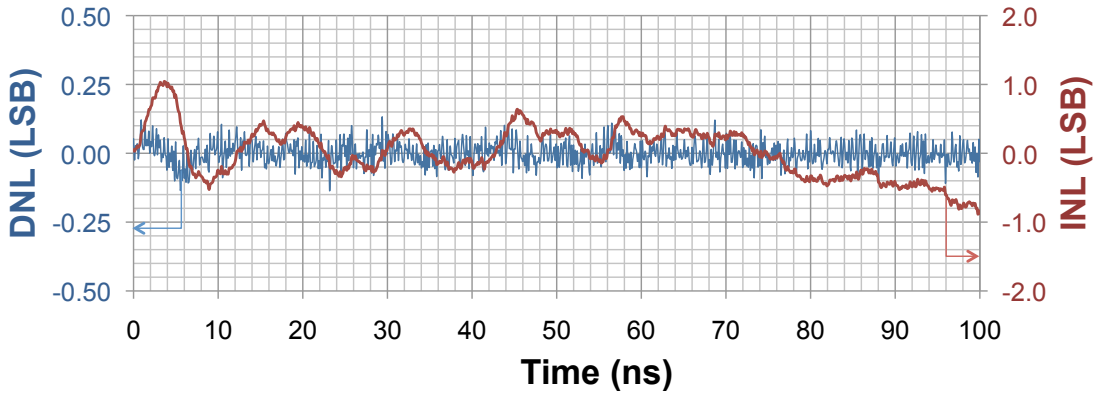
A code-density test of the SiPM plus test TI-TDC pair is performed by measuring the timing response to an uncorrelated, white-noise optical input. The code density test allows the system's timing performance to be quantified in terms of differential and integral nonlinearity (DNL/INL) [137]. For a truly monotonic converter, the DNL should not exceed $\pm$ 0.5 LSB, whilst the INL would also ideally lie within these limits. The SiPM is configured with one low DCR SPAD enabled, whose count-rate is kept below 0.5 % of the synchronisation frequency to minimise the effect of classic TCSPC pile-up, which would distort the captured histograms. To ensure statistical confidence in the results, at least 1,000 counts are captured per TDC code in all configurations. A synchronisation pulse period marginally shorter than the TDC full range is used to ensure there is no code wrap around, however this means it is difficult to measure the *full* range of the TDC in any particular setup. It should be noted that this is a system characterisation and not of the individual converters, for which this technique is more commonly used.

The test TI-TDC is initially configured to provide the finest resolution of 54 ps on the 10-bit raw test bus (see Section 4.8). A 20 MHz synchronisation pulse is used (50 ns period), allowing 91 % of the 55 ns TDC range to be tested. It is clear to see from the graph in Figure 5.14a, which shows the resulting DNL/INL of this initial test, that there are frequency artefacts in the response at $\approx$ 100 MHz and 5 GHz. The 100 MHz oscillation is caused by ringing on the *STOP* input at this synchronisation frequency (20 MHz), which is coupled onto the SPAD bias voltages with a $\pm$ 100 mV swing, thereby modulating the PDP. The higher frequency oscillation is a code probability issue within the GRO which causes every fourth code (half the beat

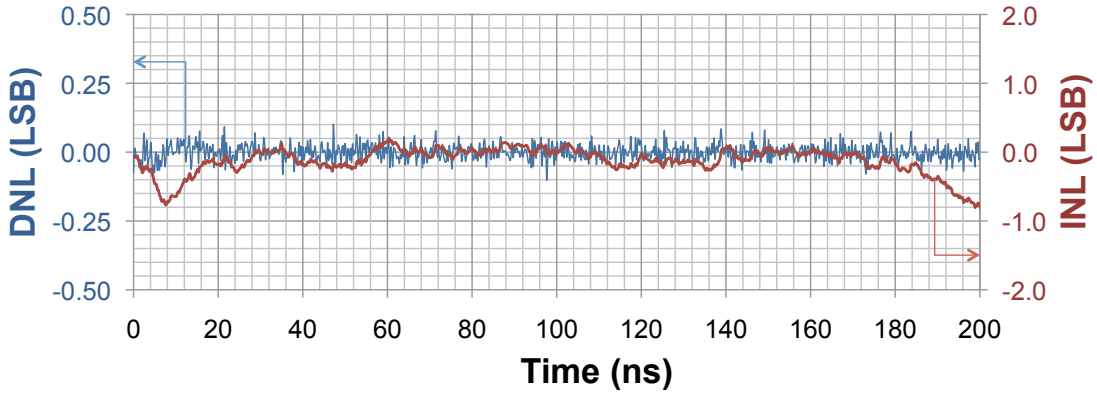
oscillation) to experience a slightly increased chance of occurrence. Despite these artefacts, the DNL is within ± 0.2 LSB, which is comfortably within the expected limits. However the INL is closer to ± 2.1 LSB, which is not ideal for fluorescence lifetime experimentation.



(a)



(b)



(c)

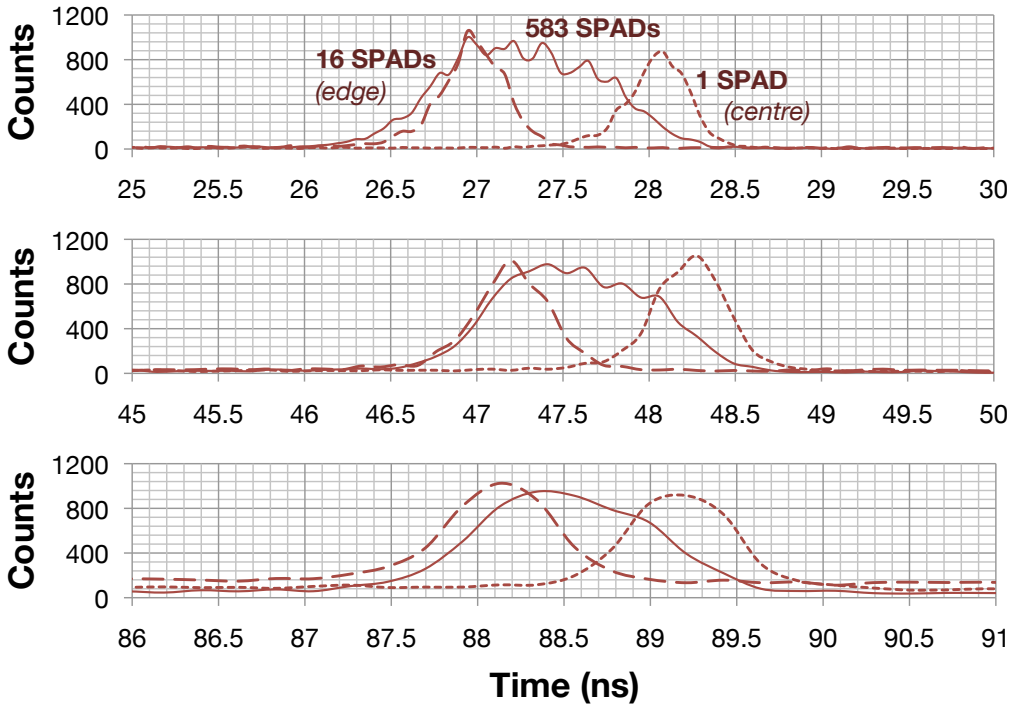
Figure 5.14: *DNL/INL code-density tests for* **(a)** *50 ns,* **(b)** *100 ns and* **(c)** *200 ns TDC ranges.*

The code density investigation continues by measuring the response with the TDCs configured to output bits 10:1 and 11:2, providing apparent resolutions of 108 ps and 216 ps, respectively. The results of DNL and INL for these experiments are shown in Figures 5.14b and 5.14c. The averaging of TDC codes by ignoring the LSBs in this way reduces (10:1) and removes (11:2) the high frequency code probability issue, providing improved DNL of $\pm$ 0.13 LSB and $\pm$ 0.10 LSB for the 108 ps and 216 ps TDC resolutions, respectively. To capture these results with the increased TDC resolution and range, the synchronisation frequency is reduced to 10 MHz and 5 MHz. At these frequencies, a disturbance caused by coupling is still apparent at the beginning of the time range, however the ringing is damped and does not oscillate noticeably beyond $\approx$ 10 ns. This results in significantly improved INL performance of $\pm$ 1.0 LSB and $+$ 0.2 / $-$ 0.8 LSB for the 108 ps and 216 ps TDC resolutions, respectively. Due to the clear linearity improvements of running the device at slower synchronisation frequencies and increased range, experiments should be performed using at least bits 10:1 of the TDC output. Although this reduces the resolution, it is deemed acceptable given the timing performance of the SPAD jitter ($\approx$ 200 ps) and TDC mismatch spread.

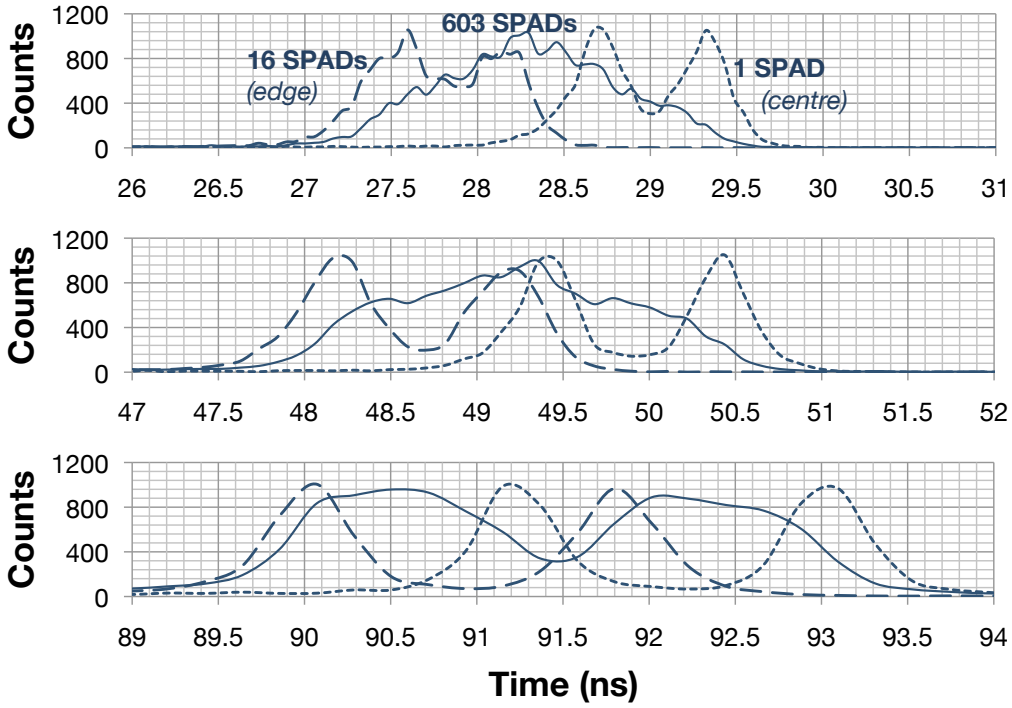
In all three graphs in Figure 5.14, the INL falls at the end of the TDC range, highlighting a gradual negative slope of the system response. This is consistent with TCSPC pile-up and is not necessarily a system issue. Due to the difficulty of minimising pile-up and maximising the number of counts per bin for statistical reasons, the experiment was not performed for higher TDC resolution/range.

5.6.3 Instrument Response Function (IRF)

The instrument (impulse) response function (IRF) of the full system is now investigated for the devices with best and worst mismatch from Section 5.4.2. The IRF is captured using different SiPM configurations and the test TI-TDC to produce histogram data. The pulsed excitation and synchronisation is achieved using a laser diode driver (Picoquant PDL-800-B) and laser head (Picoquant LDH-P) with a wavelength of 478 nm. To minimise the effect of classic TCSPC pile-up, the optical signal is attenuated to bellow 0.1 % of the excitation frequency by using the lowest power setting, leaving the beam uncollimated and placing a neutral density filter in front of the photo-sensitive area of the device. There are a number of sources of non-idealities that combine to produce the IRF: SPAD jitter, SiPM timing (routing delays, transistor mismatch), TDC jitter, TDC mismatch, synchronisation jitter and excitation jitter.



(a)



(b)

Figure 5.15: *IRFs captured using different SiPM configurations at increasing START-STOP times for* **(a)** *a low mismatch device (red) and* **(b)** *a high mismatch device (blue).*

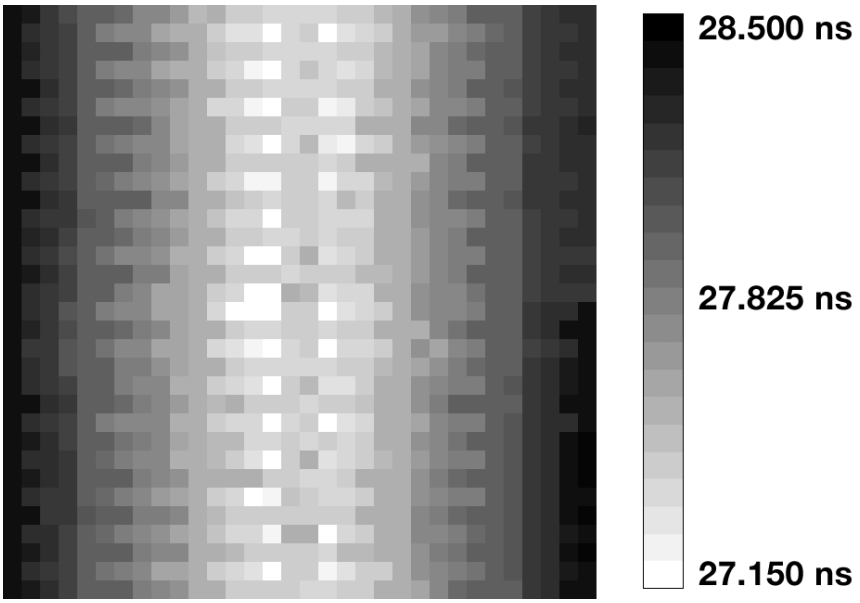
The graphs in Figure 5.15 show the results from these experiments for both devices and for increasing *START-STOP* times. A range of SiPM configurations are used: a single low DCR SPAD located in the centre of the SiPM (short dash); 16 low DCR SPADs located at the edge of the SiPM (long dash); and all SPADs with a DCR below 1 kHz, which is $\approx$ 600 in each device (solid).

For low mismatch (red), the FWHM of the IRF of one and 16 enabled SPADs is approximately equal, rising from 400 ps to 700 ps. However, their mean time varies by 1 ns in each case, meaning a time offset exists that is SPAD position dependent. This is backed up by the 583 enabled SPAD case, where the FWHM of the IRF is much wider at 1.2 ns due to the devices being located randomly across the entire SiPM array. As expected, the IRF for the device with inferior mismatch performance (blue) is significantly worse as the input *START-STOP* time is increased, rising from 900 ps to 1.3 ns for the single and 16 enabled SPAD configurations, whilst for 603 enabled SPADs, this value rises from 1 ns to 3 ns.

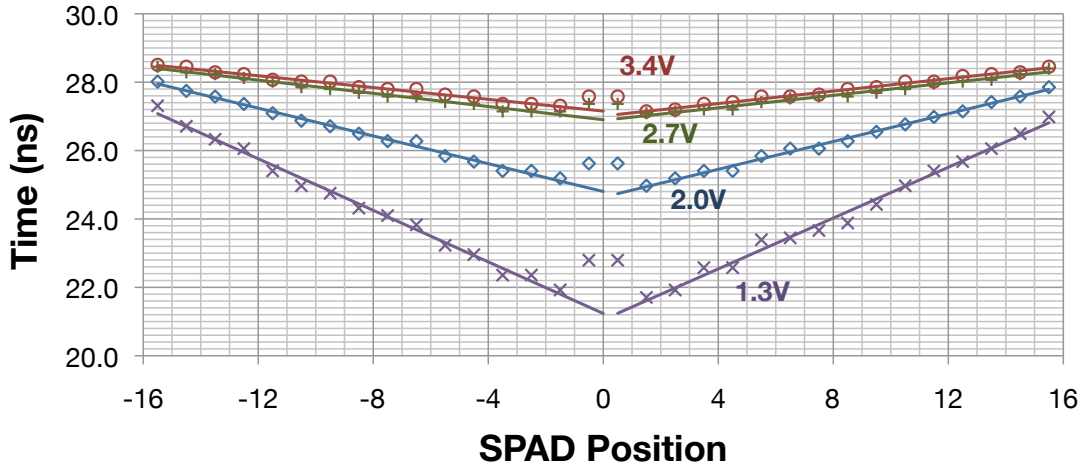
5.6.4 Position Dependent Timing

The positional dependent timing is further investigated by capturing an IRF histogram for each SPAD independently and calculating its average TDC code. Plotting this average against its position yields the image in Figure 5.16a, where an x-axis gradient from the centre to the edge is clearly evident. This gradient corresponds to the SPAD output routing, where devices in the central columns have longer routes to their level shifters than those at the edge, as described in Section 4.3. Therefore, as well as using 16 adjacent SPADs for the best fill-factor performance, the grouping also improves timing performance. A worst case difference using this device and configuration is calculated at 1.4 ns between the innermost and outermost columns.

The results in Figure 5.16b show the effect of varying V_{EB} on the SiPM timing gradient for a single row of SPADs. As can be seen, V_{EB} has a significant effect on the timing performance, with a worst-case difference of 3 ns between the innermost and outermost detectors in the row for V_{EB} = 1.3 V. This performance drop is attributed to the use of thick oxide transistors for the SPAD output buffers which do not perform optimally at these supply voltage levels, providing a reduced drive strength which is most noticeable when driving the long, high-capacitance tracks from the centre of the array. This highlights a requirement to have a high V_{EB} – preferably above 2.7 V – to achieve optimal timing performance, which is particularly important when large or spread-out groups of SPADs are used for experimentation.



(a)



(b)

Figure 5.16: *SiPM position dependent timing for* **(a)** *the full array at 2.7 V and* **(b)** *a single row at varying excess bias voltages (V_{EB}).*

5.6.5 Power Consumption

The core device power consumption is now measured using an uncorrelated (white-noise) light source and with the embedded CMM processing enabled. The results are shown in Figure 5.17 for both two- and eight-channel TI-TDCs and increasing count rate up to 650 MHz. The synchronisation frequency is set to a typical value of 10 MHz. In complete darkness, the device has a base power consumption of just below 2 mW, primarily caused by I/O and switching of the processing block. The TDCs are not operational in this case, so draw no power other than leakage. The power consumption is shown to reach 5 mW and 10 mW as the count-rate is increased towards 700 MHz, for two and eight TI-TDC channels enabled, respectively. This is likely caused by the TDCs being in constant operation at such high count rates. However, from the results presented in Section 5.3.4, the device is likely to operate below 100 MHz[2], so the typical power consumption range is 2–5 mW for the full eight-channel TI-TDC architecture with the embedded processing enabled. Combining the power consumption of the core timing and processing with that of the SiPM (500 µW for 16 enabled detectors), it is comfortably within the specification of 10 mW (see Sections 1.3 and 4.1.2).

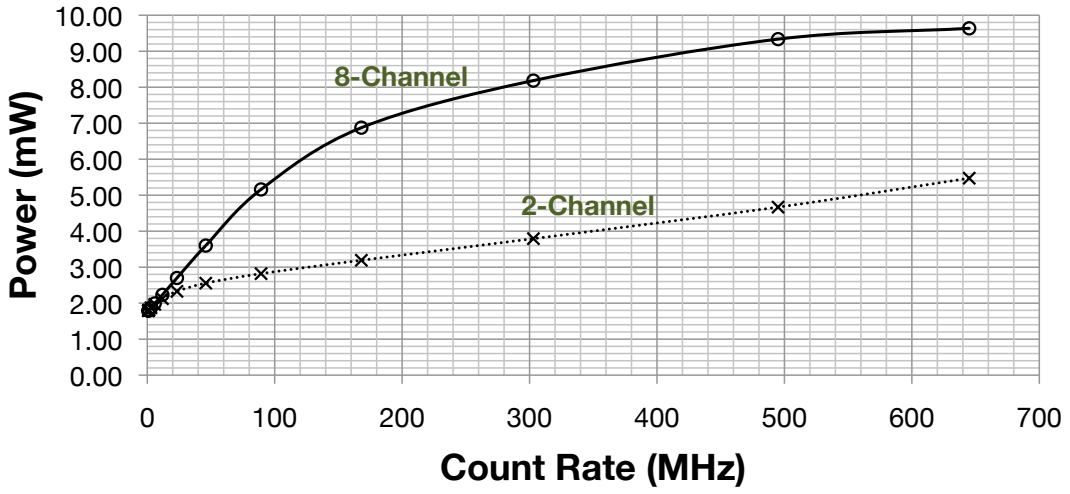


Figure 5.17: *Core timing, embedded processing and I/O power consumption.*

[2] n = 38.4 MHz @ 10 µW for 2 % photon loss

5.7 Fluorescence Lifetime Characterisation

5.7.1 Experimental Setup

The device is evaluated in bulk fluorescence experiments by mounting the hardware platform described in Section 5.2.2 onto the side camera port of a fluorescence microscope. For experimental evaluation of the sensor's performance, a range of fluorescent dyes with different lifetimes are measured using increasing laser excitation power. Although the sensor acts as a *point* detector, for experimental convenience measurements are performed as part of an existing wide-field fluorescence lifetime imaging (FLIM) system on an inverted microscope (Nikon TE2000U). The excitation source is a pulsed diode laser (Picoquant PDL-800-B/LDH-P) with a wavelength of 478 nm, coupled through the epi-fluorescence port of the microscope using a filter cube (Nikon B-2A). The laser pulse repetition rate is 5 MHz or 10 MHz depending on the lifetime being measured, and the maximum optical power reaching the back focal plane of the objective is $\approx$ 64 µW. The optical power reaching the SiPM, and hence the photon count rate, is varied by adding combinations of neutral density filters to the detection path. The excited sample volume is focussed onto the active area of the CMOS device using an additional short focal length lens. The TDC resolution is calculated prior to capturing each set of data and is 56.6 ps throughout the experiments.

The results presented within this section are those described in the accompanying publications provided in Appendix B [1–3] and use the first revision of the device which suffers from inferior DCR and afterpulsing performance[3] than has been described Section 5.3. Selecting a group of 16 adjacent SPADs with sufficiently low DCR and afterpulsing proved to be impossible. Therefore a compromise was taken to enable eight adjacent SPADs, which provide a combined DCR of $\approx$ 5.5 kHz with a 16 V bias. Using the paralysable detector theory from Section 5.3.4, this corresponds to a maximum total true count rate (n) of 8$\times$64 µW$\times$250 kHz/µW = 128 MHz, and a detected count rate (m) of 119.5 MHz with the monostable enabled, providing a 93 % efficiency in this worst case. Due to the limited time available in the microscopy laboratory, only the FLIM experiments where performed using the improved device (see Section 5.8.1) and the bulk sample experiments were not repeated. Furthermore, as the modelling in Chapter 3 proves, eight detectors is still sufficient to significantly increase the photon throughput in TCSPC, particularly for longer lifetimes, with throughputs in excess of the excitation rate being possible for only 1 % error in lifetime calculation.

[3]The inferior performance was caused by issues during device fabrication and was outwith our control.

5.7.2 TCSPC

The first experiments use the 10-bit raw TDC output bus, as presented in Section 4.8.1, to capture TCSPC histogram data for a range of fluorescent dyes. One second exposures of Rhodamine B (10 µM) and Rhodamine 6G (10 µM) in aqueous solution as well as Birch Yellow Quantum Dot in toluene (60 µM - Evident Technologies, NY, USA) are captured. These fluorophores have quoted lifetimes of 1.74 ns, 4.08 ns and 15-20 ns, respectively. To reduce the distorting effects of classic TCSPC pile-up, the count-rates of each experiment are kept below 1 % of the excitation rate. The captured histograms are shown together with the system's IRF in Figure 5.18. The FWHM of the IRF is measured as 325 ps at a mean TDC code of 25 ns and is achieved due to negligible mismatch with the test TI-TDC pair used. The histograms are curve-fitted using Edinburgh Instruments *FAST* software, providing fluorescence lifetime calculations of 1.8 ns, 3.9 ns and 15.6 ns, as shown in the figure, which are in good agreement with the quoted values. These results highlight the ability of the device to operate as a replacement for a TCSPC acquisition system, with the advantage of reducing the hardware requirements by performing the detection and timing on a single miniaturised CMOS device. Furthermore, due to the use of a TI-TDC pair, this system does not suffer from processing (conversion) dead-time pile-up, allowing an increase in photon throughput over conventional discrete component systems. However, the remaining classic TCSPC pile-up limitation of only being able to record at most one photon event per excitation period still exists.

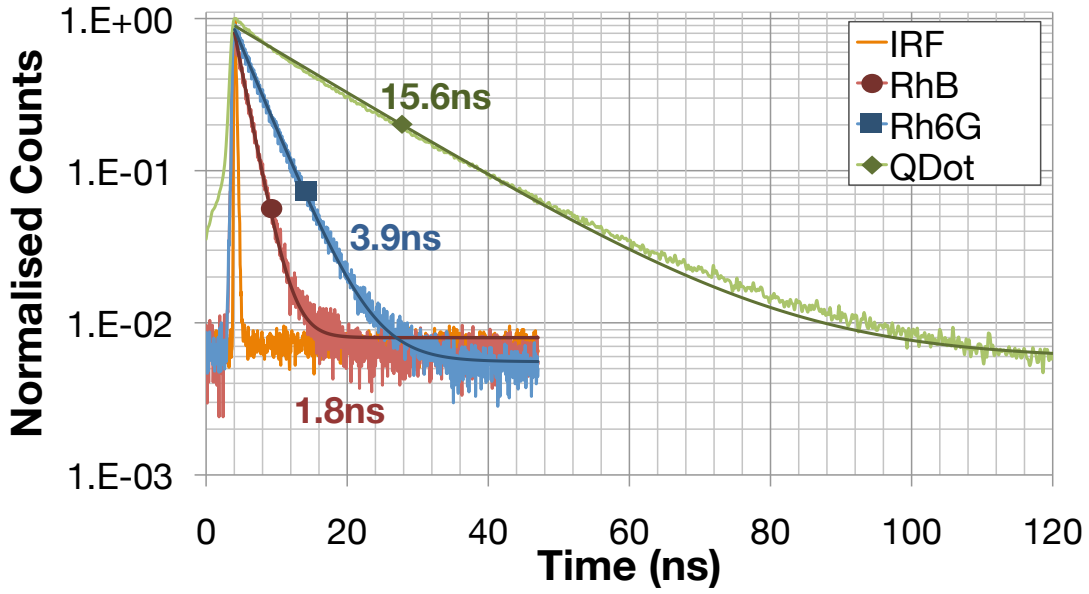


Figure 5.18: *TCSPC results for three different bulk sample fluorescent dyes.*

5.7.3 CMM

The embedded CMM pre-calculation is now enabled, using the data captured from the TCSPC experiments to configure the windowing values for *FIRST* and *LAST* (see Section 4.5.3). Initially, CMM is captured and calculated using only the test TI-TDC pair to highlight the effects of classical TCSPC pile-up. At count rates below this limit, the embedded CMM calculation successfully produces lifetime estimates of 1.7 ns, 3.9 ns and 16.5 ns after background correction (see Section 2.6.4), which are in good agreement with the quoted lifetimes and the TCSPC results presented in the previous section. Furthermore, an additional fluorophore – Rubrene in methanol – is added to the experiments, producing a lifetime calculation of 8.3 ns, which is also in good agreement with its quoted lifetime of 8.56 ns. However, as expected by simulation from Section 3.3 as shown by the filled markers in Figure 5.19 (dark line), the normalised CMM lifetime calculation falls with increasing photon throughput due to classic TCSPC pile-up.

Further background corrected CMM calculations are then performed on the same fluorophores with an increasing photon-rate for the eight-channel TI-TDC pair architecture. The normalised CMM lifetime results are shown by the unfilled markers in Figure 5.19 for all four fluorophores. The device's ability to more accurately calculate the correct lifetime value at higher photon-throughputs is clearly apparent, despite using only eight detectors. In all cases, a photon throughput equal to the excitation rate is demonstrated for a worst case error of 4 % for the shortest lifetime (Rhodamine B) and a best case of only 1 % for the longer lifetime fluorophores (Rubrene and Quantum Dots). Furthermore, a photon throughput of five times the excitation frequency is possible for the Quantum Dot sample for a 5 % error in calculation, in contrast to the single channel error of $\approx$ 60 % at the same photon rate. All of these results are consistent with the expected performance defined in Figure 3.24.

The known device and experimental variables can be used to further validate and increase confidence in the model developed in Chapter 3. These variables include the number of SPADs (N_D = 8), SPAD dead-time ($t_D \approx$ 10 ns), DCR (5.5 kHz), SiPM pulse width (t_P = 540 ps), number of TI-TDC timing channels (N_T = 8) as well as the fluorescence lifetimes and excitation frequencies used in each experiment. The similarities between the model and laboratory results are clearly shown at the bottom of Figure 5.19 by the solid coloured curves. The slight discrepancies for the shorter lifetime fluorophores is attributed to the sources of system non-idealities that cannot be measured, such as TDC mismatch.

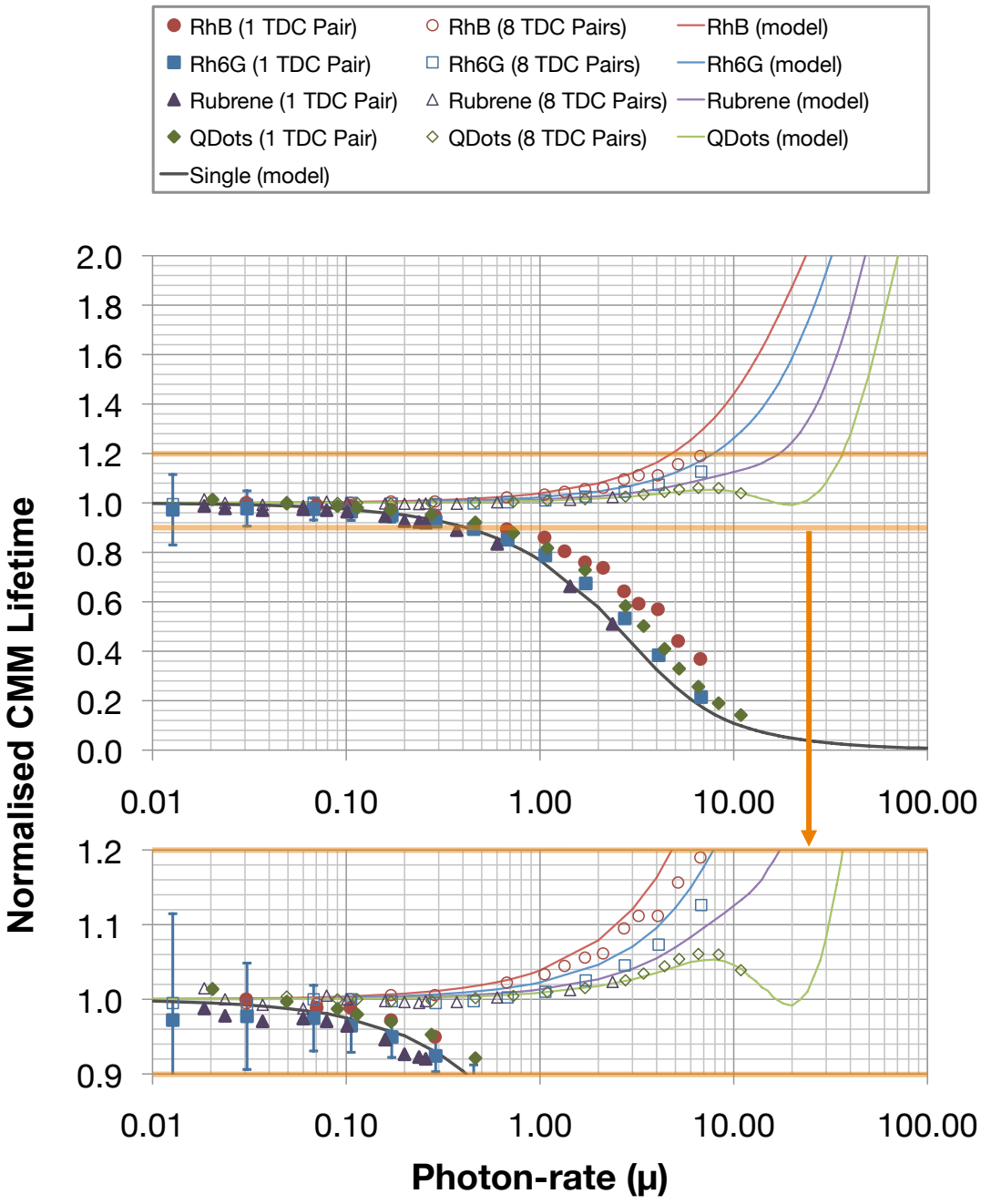


Figure 5.19: *Effect of increasing the photon-rate (μ) on the normalised CMM calculation for Rhodamine B (red), Rhodamine 6G (blue), Rubrene (purple) and Quantum dots (green) using one TI-TDC pair (filled markers), eight TI-TDC pairs (unfilled markers) and simulation (solid curves).*

The results presented in Figure 5.19 are captured and calculated by combining data from 1,000×1 ms exposures, resulting in a total exposure time of one second. Using these multiple short exposures, analysis of the precision achieved with the implemented device is shown for Rhodamine 6G in Figure 5.20 by the blue markers. In this instance the precision is graphed in terms of photon-rate (μ), where the total number of photons $N_C \approx \frac{\mu \times f_E}{1000} = \mu \times 10,000$. This relationship does not hold true at higher count rates due to the number of photons lost to the various forms of pile-up.

As expected, the results using both a single TI-TDC pair (×) and eight TI-TDC pairs (○) are generally in line with the data presented in Sections 2.6.5 and 3.9, which use real single-channel data and modelled data, respectively. Furthermore, re-running the model from Chapter 3 using the actual experimental variables produces very similar results, as shown by the solid curve in Figure 5.20. The slight disparity between the model and captured results at high photon rates can be attributed to system non-idealities that are not included in the model, such as an increase in noise due to uncorrelated background light at high laser intensities.

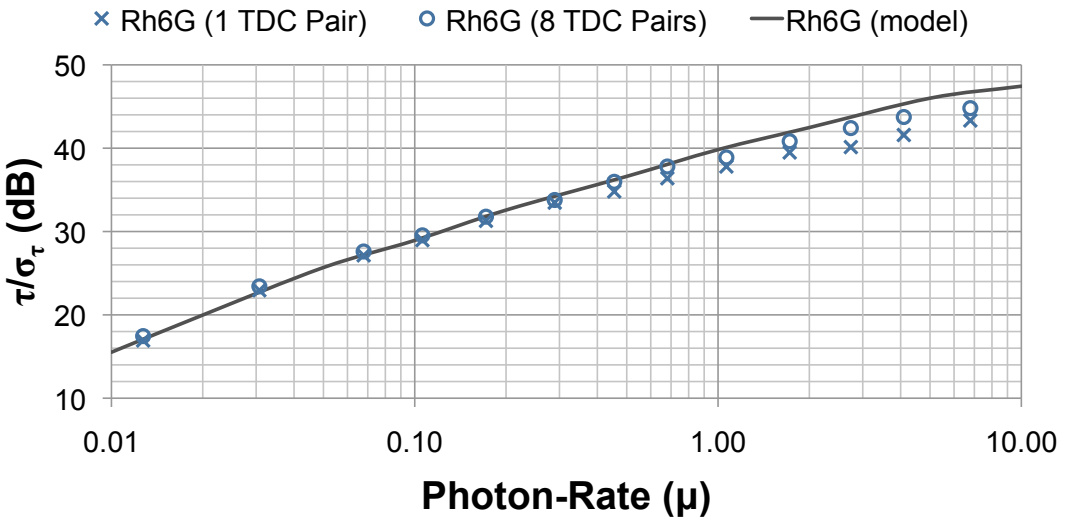


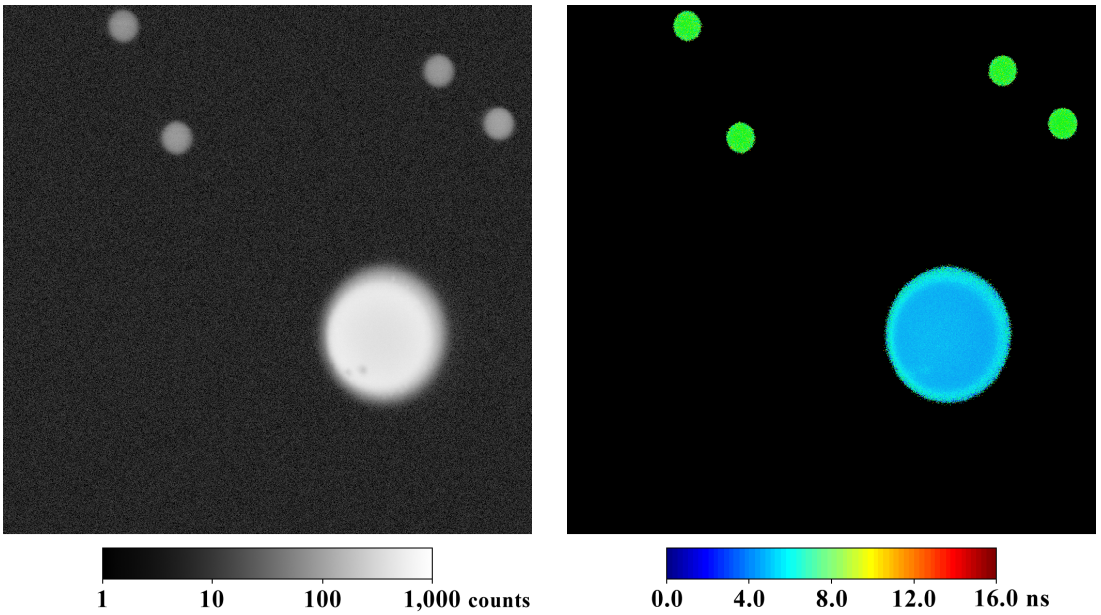
Figure 5.20: *Effect of increasing the photon-rate (μ) on the precision of the CMM calculation for Rhodamine 6G using one TI-TDC pair (×), eight TI-TDC pairs (○) and simulation (solid line).*

5.8 Fluorescence Lifetime Applications

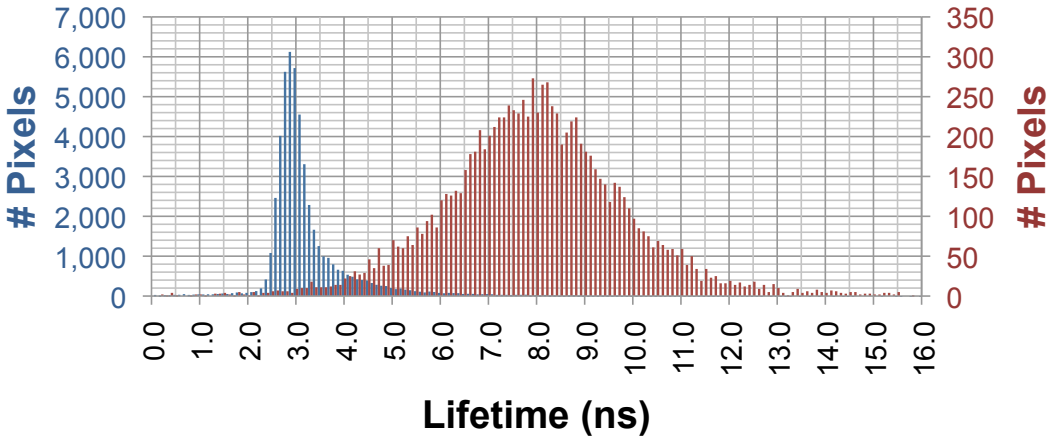
5.8.1 Fluorescence Lifetime Imaging (FLIM)

The microscopy set up is now combined with a custom scanning head and a pin-hole is introduced into the fluorescence emission path to perform confocal scanned fluorescence lifetime imaging (FLIM). The device used for the experiments reported in this section is the improved DCR part, which was necessary to achieve the required signal to noise ratio (SNR) performance. The scanner and its control software offers flexibility in choosing image resolution, pixel dwell time and single- or multiple-pass scans. However, in this work it is only configured to perform a single-pass scan. To optimise the optical set up and the system's FPGA firmware and software for efficient data capture, multiple test images of fluorescent blocks are acquired with image resolutions and pixel dwell times ranging from 160×160 to 1024×1024 and 40 µs to 200 µs, respectively.

The images in Figures 5.21a and 5.21b show log-intensity and CMM-FLIM images, respectively, of a commercial yellow-green fluorescent 15 µm diameter bead (Invitrogen Detection Technologies, UK) with a lifetime of $\approx$ 3 ns, mixed with unknown fluorescent 5 µm diameter beads with lifetimes of $\approx$ 8 ns. The beads are chosen due to their lack of photobleaching, as well as their differences in intensity. The images have a resolution of 1024×1024 (1 Megapixel) and are captured with a 100 µs pixel dwell time, resulting in a $\approx$ 100 s acquisition. The intensity image is plotted on a log scale due to the largely varying intensities of 500 kHz and 5 MHz for the small and large beads, respectively. The experiment uses a 10 MHz excitation repetition rate, so the brighter bead is at least five times beyond the pile-up limit. The CMM-FLIM image is thresholded to display black for pixels whose counts are below the noise floor, leaving only the beads visible. The CMM pixel values are background corrected and plotted on a false colour scale between 0.0 ns (blue) and 16.0 ns (red). The extracted lifetimes of the two beads appear to lie around their expected values of 3 ns and 8 ns; to confirm this, pixel histograms are produced, as shown in Figure 5.21c. Due to the much higher number of pixels of the larger bead, the image is split into two regions (top and bottom) with their corresponding histograms shown in the figure by the red and blue bars. The centres of the approximately Gaussian distributed histograms lie at $\approx$ 2.9 ns and 8.0 ns, which are in good agreement with expectations. The differing distribution spreads are a consequence of the relative intensity difference between the two samples, with the dim sample only providing $\approx$ 50 counts/pixel for the CMM calculation.



(a) (b)



(c)

Figure 5.21: **(a)** *Log intensity,* **(b)** *thresholded background corrected CMM and* **(c)** *Histogram of FLIM showing top half (red) and bottom half (blue).*

These results not only highlight the successful integration of the device and its system into a scanned confocal FLIM set up, but it also demonstrates the ability to operate beyond the pile-up limit. Moreover, the ability to simultaneously capture dim and bright samples – without the need to reduce excitation power for the bright sample and hence increase the exposure for the dim sample – emphasises impressive dynamic range capability and ease of use.

5.8.2 Simulated Flow

Finally, a *simulated* flow experiment is performed using a mixture of two types of fluorescently labeled polystyrene beads. Commercial yellow-green fluorescent 2.4 µm diameter beads (Interfacial Dynamics Corporation, Portland, OR) and 2 µm diameter beads labeled with fluorescein have distinct lifetimes of ≈ 7 ns and 2.7 ns, respectively. Due to the unavailability of a suitable flow setup and the complexities of building a custom one for this experiment, the *flow* is simulated by manually scanning the mixed sample through the laser focus. Many sequential CMM calculation exposures of 4 ms were acquired to match the relatively slow manual scanning process. The results, as shown in Figure 5.22, clearly show the device's ability to distinguish these two labeled beads, with their expected lifetimes marked by the horizontal lines. This is despite the high DCR caused by the device, which is clearly apparent in the bottom graph. An intensity threshold is applied to the CMM calculation, marked on the bottom graph by the green line. The brighter of the two samples produced a photon count rate of 2.5 MHz, a photon-rate of 25 % of the 10 MHz excitation rate, highlighting the device's resistance to TCSPC pile-up. If this experiment was performed using a typical single-channel TCSPC setup with discrete components, the laser power would be tuned to provide a photon-rate of 1-5 % of the excitation rate for the brightest sample, significantly reducing the photon throughput of both samples by a factor of 5-25.

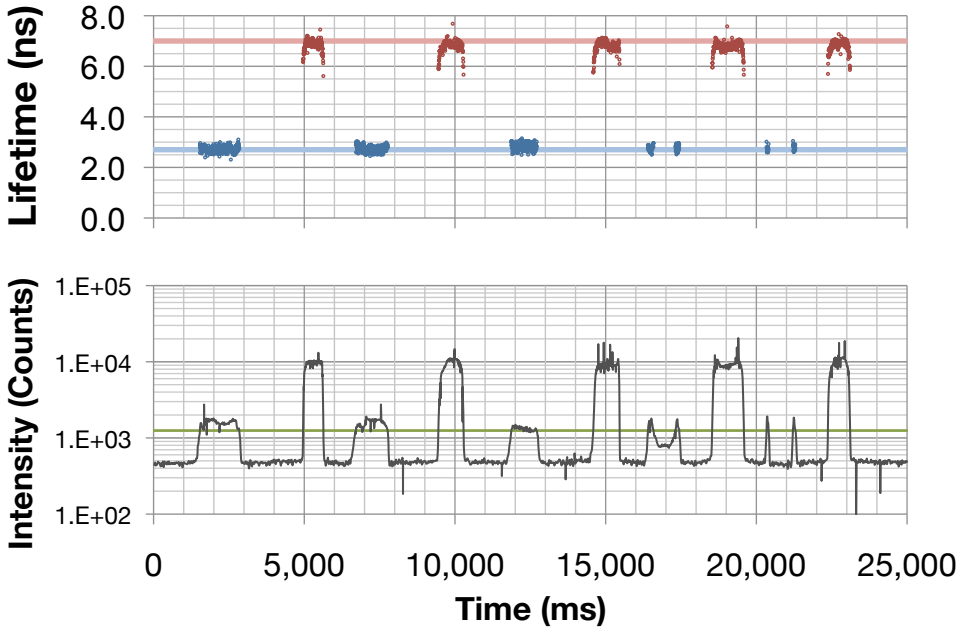


Figure 5.22: *Results from simulated flow experiment of a mixture of fluorescent beads.*

5.9 Conclusions

This chapter has described the supporting development platform required to not only test and characterise the device, but to use it in practical fluorescence lifetime experiments. The platform consists of a PCB which contains sockets for the sensor and an off-the-shelf plug-in FPGA board, the latter of which required custom firmware to drive and process data to and from the device. A custom software application is also developed to configure the FPGA and device and to further process and visualise data captured from it. Furthermore, many routines are written in the software to semi-automate specific characterisation tasks, such as DCR calibration.

The characterisation of the device began by testing the individual components separately. The SiPM presents impressive throughput performance, allowing photon count rates of up to 50 MHz and 700 MHz with the monostable disabled and enabled, respectively. Although DCR performance does not meet the initial expectations (see Figure A.1), it is still possible to select a region of 16 adjacent SPADs for experimentation which produce a combined DCR of below 2 kHz. Characterisation of the TDCs as individual components is in line with the expectations from [12], however the expected mismatch issue is highlighted by a worst case TI-TDC resolution variation on a single device of 1.0 ps (52.2 ps and 53.2 ps). This has the effect of creating a large FWHM of the TI-TDC pair of over 1 ns at a timing delay of 50 ns. The linearity of the 1.43 ns resolution delay line is promising and its inclusion in the device proved invaluable for experimental flexibility.

System characterisation is then performed by testing the SiPM and timing architecture together. Measuring timing linearity using code-density tests proved problematic due to a limitation of using the SiPM as the white noise source, where an oscillation on the bias voltages significantly affected INL. Fortunately this issue is minimised by reducing the excitation frequency below 20 MHz and increasing the TDC range accordingly. Measurements of the optical IRF are device dependent due to TDC mismatch, providing FWHMs of $\approx$ 500 ps and $\approx$ 1.3 ns for the best and worst devices, respectively, at 50 ns delay with 16 adjacent SPADs enabled. This IRF result is not only caused by SPAD and TDC jitter, but the SiPM also presented a positional timing variation caused by routing delays between the SPADs and their monostables, which is longer from the central columns than the outer columns. The worst case variation is minimised to below 1 ns by ensuring $V_{EB} > 2.7$ V. Finally, power consumption is measured to be between 2–5 µW at typical operating count rates, which is within the device specification.

Results from TCSPC experimentation successfully demonstrate the ability of the system to capture histogram data for post-processing with zero timing dead time, highlighting the suitability of replacing the expensive, complex and power hungry discrete component TCSPC set up. Enabling the CMM calculation then opens the way for *real-time* fluorescence lifetime processing for suitable applications. The most promising result however is the increased throughput made possible by the multiple-channel timing architecture. Despite the use of an inferior DCR part, throughputs in excess of the excitation frequency are shown to be possible for minimal error in lifetime calculation of < 4 %, whilst for the longest lifetime fluorophore (≈ 16.5 ns), a throughput of five times the excitation frequency for a 5 % error in lifetime calculation is possible. The similarities between these results and using the model from Chapter 3 with the same parameter set are very promising, giving confidence of further performance gains with the improved DCR parts.

The chapter finishes by performing trials of two practical fluorescence lifetime applications: fluorescence lifetime imaging (FLIM) and *simulated* flow based sorting. Both experiments were successful, producing accurate CMM lifetime extraction measurements of fluorophores producing photon count rates both within and beyond the classic TCSPC pile-up limit. These results highlight the biggest advantage of the developed system; the ability to simultaneously capture lifetime data of samples whose intensities are over an order of magnitude apart without the need to reduce excitation power and therefore increase exposure time. Not only does this highlight an impressive dynamic range capability, but it demonstrates an ease of use not commonly associated with TCSPC experimentation.

6
CONCLUSIONS

6.1 Summary

The research presented within this thesis successfully demonstrates the design, implementation and operation of a miniaturised optical sensor with embedded processing for time-domain fluorescence lifetime experimentation using time-correlated single photon counting (TCSPC). The manufactured device is shown to be capable of operating beyond the classic TCSPC pile-up limit of typical discrete component arrangements, whilst providing fluorescence lifetime calculation estimates in *real-time*. This has been achieved by understanding the limitations of current state of the art TCSPC instrumentation and investigating applicable CMOS technologies and architectures capable of overcoming these limitations. Detailed modelling of a proposed architecture leads to the specification of a device that is designed, simulated and manufactured in an advanced 130 nm standard CMOS imaging process. Finally, after thorough characterisation of its individual components, the device is used to perform practical fluorescence lifetime experiments, where it is shown to successfully meet the original project aims, achieving photon count rates more than an order of magnitude greater than the TCSPC pile-up limit.

A critical review of state of the art single and multiple channel TCSPC architectures is presented in Chapter 2, with a focus on hardware and software techniques currently used to overcome or minimise the effects of pile-up. Following this, a review of state of the art CMOS technologies currently available to perform single photon detection and picosecond resolution timing is described, before techniques to perform embedded fluorescence lifetime calculations are presented. These investigations direct the choice of an *integrated* CMOS architecture using single photon avalanche diodes (SPADs) in a silicon photomultiplier (SiPM) arrangement with temporal compression, combined with a multiple channel time-interleaved time-to-digital converter (TI-TDC) array, aimed at providing a more efficient and easy to use technique to overcome TCSPC pile-up. A pre-calculation of the 100 % photon efficient high throughput centre-of-mass method (CMM) is chosen as the embedded calculation technique.

A MATLAB model is developed in Chapter 3 to investigate the performance of the chosen integrated architecture with different device parameters and under different experimental conditions. The performance is quantified both in terms of the ability to reduce the number of photons lost due to pile up and the maximum photon throughput available for a given loss of accuracy in lifetime calculation using CMM. The investigation concludes that a 16 detector SiPM and an eight channel TI-TDC pair timing architecture is capable of providing a photon throughput equal to or above the excitation frequency for a 1 % error in calculation of the lifetime. This performance is possible so long as the SiPM output pulse-width is at least ten times shorter than the fluorescence lifetime being measured. Finally, timing mismatch between the multiple TDCs is added to the model and a technique to minimise its effect by randomising the TDC utilisation and performing a calibration, is described.

The design of a CMOS device based on the chosen architecture and parameters from Chapters 2 and 3, respectively, is presented in Chapter 4. The use of an advanced 130 nm standard CMOS technology allows the typically large and expensive discrete components required for a TCSPC experiment to be integrated onto a single miniaturised device. Furthermore, implementation in a standard CMOS process offers low manufacturing costs for high volume production. A 32×32 element SiPM array is designed to make efficient use of the available silicon area, whilst providing the experimental flexibility to choose a suitable region of 16 low dark count rate (DCR) detectors. A compressed SPAD output pulse-width of 250 ps is chosen based on the results of extracted circuit simulations. A token-passing circuit is developed to distribute photon events from the SiPM to the multiple channel TI-TDC architecture. However, due to metastability issues within this circuit, it could not be designed to minimise the effect of TDC mismatch by randomising the distribution. To facilitate test and characterisation of the device, it is also designed to operate as an integrated single-channel TCSPC sensor.

The device is tested, characterised and demonstrated in practical fluorescence lifetime experiments in Chapter 5. A higher than expected DCR distribution is measured, with only 30 % of SPADs being under 100 Hz. Timing performance is shown to be optimal by reducing the apparent TDC resolution from 54 ps to 108 ps or 216 ps by averaging its output codes, with the major remaining source of error coming from TDC mismatch, as expected. The device is successfully demonstrated in bulk sample fluorescence lifetime experiments by capturing single-channel TCSPC and CMM data, however these results are still limited by pile-up at photon rates above 10 % of the excitation frequency. Enabling the eight-channel TI-TDC

architecture with CMM allows photon throughputs in excess of the excitation frequency for a 4 % error in calculation, with rates of up to five times the excitation frequency being possible for the same error with the longest lifetime fluorophore (16.5 ns). Finally, the device is used in practical proof of concept fluorescence lifetime imaging (FLIM) and simulated flow experiments, demonstrating its ability to provide an increased dynamic range when simultaneously measuring bright and dim fluorescent samples.

6.2 Critical Discussion

The results captured using the developed prototype fluorescence lifetime sensor demonstrate that photon throughput rates in excess of the classic TCSPC pile-up limit are possible by integrating single photon detection, picosecond timing and embedded signal processing on a single CMOS substrate. The device and its supporting evaluation platform (PCB hardware, FPGA firmware and software) progresses the state of the art in TCSPC instrumentation, providing a higher throughput, miniaturised, lower-cost, lower-power and easier to operate alternative to traditional fluorescence lifetime experimentation techniques. The inclusion of a test mode to provide raw timing data has proved invaluable during experimentation and is in its own right a significant advance in TCSPC instrumentation, for the same reasons as above. However, as expected from a first iteration prototype, there are a number of shortcomings and areas for improvement, as will be presented below in this critical discussion of the design, implementation and results.

The silicon photomultiplier (SiPM) implementation has a number of minor drawbacks that result in non-idealities in the captured results. Although the findings from Chapter 3 specified that 16 detectors where sufficient to achieve the specified throughput gains, a 32×32 SiPM was implemented to maximise silicon area utilisation. Due to DCR issues with early versions of the chip, the flexibility offered by this larger number of detectors proved extremely useful for finding a compact region of suitably low DCR SPADs for experimentation. However, the larger format SiPM exposed a position dependent timing variation that adds an additional source of error to the instrument's timing response (see Section 5.6.4). Fortunately this is minimised by the careful selection of SPADs and by supplying a high V_{EB} to increase the output inverter drive strength. Furthermore, the SiPM size is the limiting factor for the compressed output pulse-width, which is critical to achieving improved throughput performance. Techniques to reduce this pulse-width, including decreasing the SiPM size, are presented in Section 6.4.2.

The high-speed asynchronous nature of incoming photon events from the SiPM to the event distribution router circuit presented a metastability issue with the original *free-running* design of this block, which resulted in the *token* being lost until reset at the beginning of the next exposure (see Section 4.4.3). Unfortunately due to time limitations during the chip design phase of the project, a compromise was taken to revert to a *resetting* router which is capable of correcting for metastability by retaining the *token* at all previous states. This approach meant that there was greater timing variability between devices caused by TDC mismatch. A technique to overcome both the metastability and the timing performance loss of a *resetting* router is presented in Section 6.4.3.

The embedded processing block consists of a pre-calculation of the 100 % photon efficient single-exponential centre-of-mass method (CMM). This pre-calculation, where the final division is performed off-chip, is necessary to reduce the bandwidth requirements by providing data compression when it is operated at high photon throughputs. Partitioning the calculation in this way was the most efficient use of resource given the time constraints of the project. Developing the hardware to also perform the division on-chip is the next logical step to improve the implementation, however doing so would incur a significant area increase and would require more complex control. Extensions to the basic CMM calculation (see Section 2.6.4) are the focus of ongoing parallel research and include the development of hardware efficient techniques to perform background correction [134] and range extension, which allows lifetimes to be computed when $T < 7 \cdot \tau$ [10]. Furthermore, although the single exponential model is effective at *contrasting* different fluorophores – which is useful for high-rate diagnostic applications such as flow based sorting – a hardware efficient approach to calculating two-exponential decay parameters has recently been developed [6], which is important for applications such as FLIM-FRET. However, as with performing division on-chip, each of these extensions has its own complexities and area requirements. Despite all of the advantages of *real-time* fluorescence lifetime calculation techniques, capturing the raw TCSPC data at photon-rates above the pile-up limit is also of great interest, so techniques to achieve this are presented in Section 6.4.4.

The first fabricated device, which was used to capture the initial TCSPC and CMM characterisation results (see Section 5.7), suffered from higher than expected DCR performance. This limited the number of SPADs that could be enabled for experimentation; a group of 8 was used, providing a total DCR of 5.5 kHz. Furthermore, the excitation

synchronisation pulse created a ringing disturbance throughout the chip, including being coupled onto the SPAD bias voltages, as described in Section 5.6.2. This has the effect of modulating the SPAD PDP synchronously with the excitation, creating a serious timing non-linearity. Fortunately, by reducing the excitation frequency from 20 MHz to 10 MHz or below, the oscillation is damped and linearity is improved for the majority of the TDC range. Despite these difficulties, the device is still able to demonstrate a significant improvement over conventional pile-up limited approaches, presenting throughputs in excess of the excitation frequency for minimal calculation error (< 4 %). The similarities between these captured results and the results from simulations using the same device and experiment parameters, as shown in Figure 5.19, is very encouraging as it provides confidence that with the improved DCR parts, performance in line with the expectations from Chapter 3 are entirely possible.

The scanned fluorescence lifetime imaging (FLIM) and simulated flow results presented in Section 5.8 highlight the suitability of the device – and its accompanying hardware and software control – to perform real-world TCSPC applications. The ability in scanned FLIM to simultaneously measure fluorescence lifetime values of areas with contrasting high and low brightness – without compromising the laser/excitation intensity, which would increase exposure time – demonstrates impressive dynamic range performance and ease of use. All of the results presented in this thesis show that the current implementation of the device is best suited to measuring longer lifetime fluorophores (> 5 ns) due to the channel pile-up limitation caused by the SiPM output pulse-width. This is an important and useful advantage as longer lifetimes are currently most limited by classic TCSPC pile-up due to the relationship with the reduced excitation frequency required to fully resolve the lifetime decay. However, reducing the SiPM pulse-width further, or negating channel pile-up completely will make a multiple timing-channel SiPM architecture even more powerful.

6.3 Future Work

There are a number or areas of work that could be undertaken using the current iteration of the device, if sufficient time and resource was available. Most importantly, the TCSPC and CMM characterisation experiments should be performed with 16 enabled detectors using an improved DCR part. This would ideally confirm the expectations from simulation and provide a better picture of the performance gains possible using the multiple timing channel SiPM architecture. Additionally, although the simulated flow experiments worked well as an initial proof of

concept, integrating the device into real flow sorting apparatus would provide evidence of the significant advances possible by combining the application with this new TCSPC technology. Achieving these advances also requires the implementation of *real-time* CMM background correction and division on FPGA, rather than the approach of software post-processing used for the characterisation results presented in this thesis. Finally, combining the sensor with a miniaturised pulsed optical source – such as micro-LEDs [34] or laser diodes [139] – to create a microsystem, would further miniaturise experimental setup for flow experiments, as well as open the door for other interesting applications of miniaturised fluorescence lifetime systems, such as explosives sensing [140], where work has already begun.

Before the completion of this work, a further iteration of the device was implemented and manufactured, replacing the blue sensitive SPAD with a deep-junction variant which has significantly improved near-infrared performance, providing a peak PDE of 44 % at 690 nm and 20 % at 850 nm [141]. The integration of this new SPAD not only provides an optimal wavelength choice for fluorescence lifetime applications with a longer wavelength emission, but it also enables the use of the sensor in other important application areas such as ranging [142] – which requires non-visible light – and time domain functional near-infrared spectroscopy (fNIRS) [143]. The use of the device for the multiple detector fNIRS application would also benefit greatly from bring-up and testing of the networking feature, which could not be performed due to the time, cost and complexity of producing a second custom PCB. Finally, although the device itself is not directly applicable, the concepts and technology developed throughout this research can also be used to make advances in additional application areas, such as positron emission tomography (PET) [144] and fluorescence lifetime endoscopy [145]. Finally, micro-lensing of the device is currently being performed to improve fill-factor of the photo-sensitive area, which will bring advances to all application areas mentioned above.

6.4 System Improvements

6.4.1 System Architecture

The current limiting factor of the system for further throughput improvements, particularly for shorter ($<$ 5 ns) lifetime fluorophores, is the finite pulse-width through the single SiPM output channel. Although smart design techniques can be used to reduce the pulse-width, it will always be process limited. Therefore the only solution to provide further throughput

improvements is to increase the number of channels beyond one ($N_C > 1$). This leads to a new architecture proposal for future iterations of the device, as shown in Figure 6.1, where the multiple channels create multiple *sub*-SiPMs ($N_S > 1$). However, the SiPM design of this proposal is complicated as the independent *sub*-SiPMs must be spatially interleaved so that when selecting a region of adjacent detectors for experimentation, the channel usage is balanced. A suite of new simulations is required using an updated architectural model, so that this new proposal can be understood and an informed decision made on the optimal parameters, particularly for the number of channels necessary to achieve a specified throughput increase.

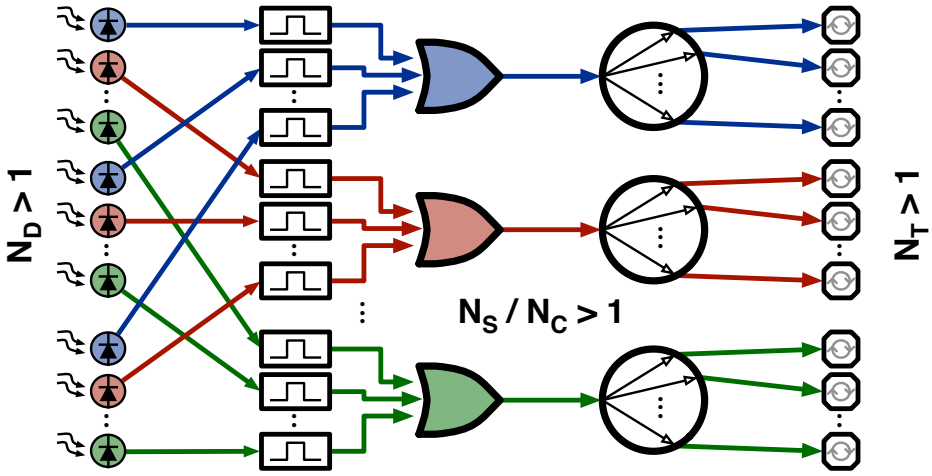


Figure 6.1: *Proposed spatially interleaved sub-SiPM system architecture.*

6.4.2 SiPM

With only 16 individual detectors required for experimentation, the first possible area for improvement of the SiPM is a reduction in size to 16×16 or even 8×8. This would significantly improve both the minimum available channel pulse-width and the position dependent timing delay error, whilst still providing experimental flexibility to avoid high DCR SPADs. To further improve the positional timing error, modifications should be made to the routing between SPADs and the OR-tree to equalise the track distance and capacitance. Now that the work in [112] has been proven in silicon and successfully demonstrated on the bench, techniques such as using a *honeycomb* structure, sharing guard-rings and placing logic circuitry in the periphery can be used to achieve fill-factors in excess of 50 %. As discussed in the previous sub-section, any new SiPM architecture should consist of multiple spatially interleaved *sub*-SiPMs to provide more than one output channel to the timing circuitry.

The conceptual diagram in Figure 6.2 shows a possible implementation of the proposals presented above, in the form of an 8×8 honeycomb structured SiPM. Four independent and spatially interleaved *sub*-SiPM channels are denoted by the colours (red, green, blue and purple), with each channel requiring its own OR-tree. The region highlighted in orange is an example of 16 detectors enabled for experimentation, where there are an equal number of SPADs belonging to each independent channel. As each *sub*-SiPM is smaller than the full SiPM (which itself is smaller in size), the OR-tree depth is significantly reduced, allowing the use of shorter monostable generated SPAD pulse-widths. The numbers and letters in the figure map the SPADs to their first level OR gate, with all the paths equalised to five hexagon edges to balance routing delays[1]. In practice, the OR-tree will not be as large as shown in the figure, as the second and subsequent levels can be physically *folded* back into gaps in the first level, due to the small area of the logic cells. Reducing the area of the SiPM in this way also provides additional space on the silicon for improved signal processing, such as background correction and division of the CMM calculation.

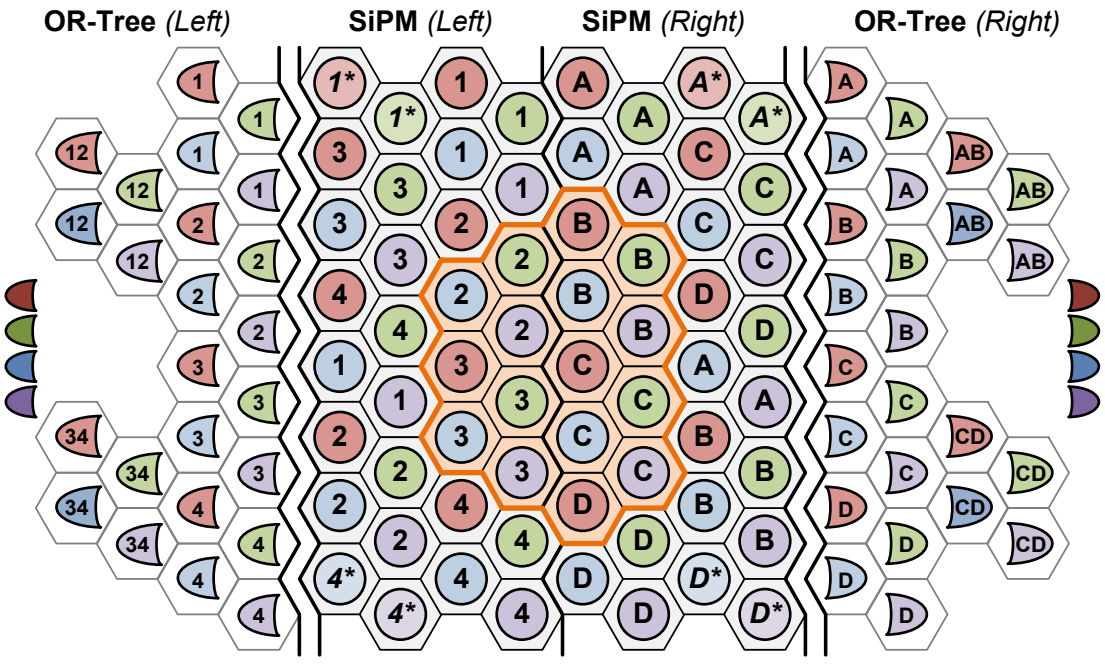


Figure 6.2: *Proposed 8×8 honeycomb SiPM architecture with four spatially interleaved channels.*

[1] With the exception of those marked n^*, which require additional (*functionally* unnecessary) routing to be added to their paths.

6.4.3 Timing

The major source of timing error within the system is TDC mismatch. There are two relatively simple modifications that can be made to the design to improve performance in this area. The first technique requires the minimum amount of change to the current timing architecture, so would be the fastest and easiest to implement. The mismatch error is emphasised by the varying position of the histogram peak, which is caused by the commonly adopted *reverse START-STOP* timing approach used for TCSPC. Combining a *forward* timing mode with the successful embedded delay line would allow the decay peak to be positioned close to time zero ($t = 0$) and hence minimise the effect of gain mismatch at the point where most photons arrive. The results from running the CMM mismatch simulations from Section 3.10 using a *forward* timing mode and *resetting* router are shown by the blue error bars in Figure 6.3. The worst case errors caused by mismatch are much smaller than for the *reverse* timing mode with a *resetting* router (grey error bars) and are comparable to the *free-running* router (see Appendix A.5). Using a *forward* timing mode has the drawback of producing an increased power consumption that is *inversely* proportional to photon activity[2]. However, this is not expected to be a major issue given the relatively low power consumption compared to a discrete component TCSPC arrangement. The second modification is to introduce a pseudo-random linear feedback shift register (LFSR) element into the routing circuitry to randomise the reset state and hence TDC usage. This would have the effect of making the safer *resetting* router operate like a *free-running* router. Combining this technique with a *forward* timing mode yields further performance improvements, as shown by the red error bars in Figure 6.3.

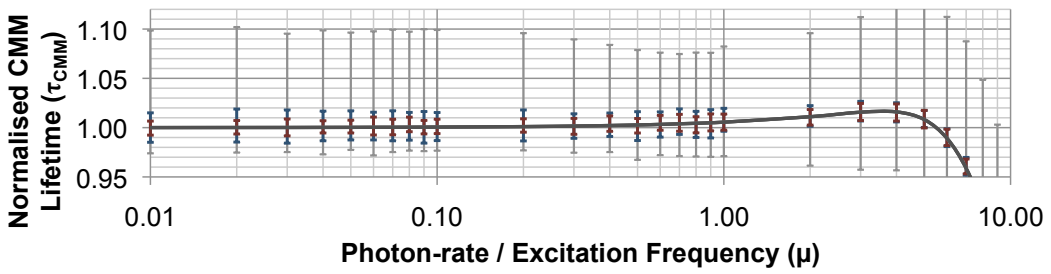


Figure 6.3: *Worst case CMM calculation errors (error bars) from 100 random TDC mismatch configurations using a forward timing mode with resetting (blue) and randomised (red) routers.*

[2]The TDCs will all run continuously when there are no photons and will be stopped earlier (on average) as the number of detected photons is increased.

6.4.4 Data Processing & Acquisition

In addition to the CMM calculation enhancements introduced in Section 6.2, there is scope to optimise the existing embedded processing and provide access to raw TCSPC data for multiple photon events per excitation period. The hardware requirements for the CMM calculation can be significantly simplified by using the TDC itself to perform the code accumulation by not resetting it between events. This would not only save the area required by the accumulating circuitry, but would negate the need for pairs of TDCs, as there would be no reset dead-time. To achieve this, a *reverse* timing mode is necessary as the TDC should only be *re*started when triggered by a photon event. The major downside to a non-resetting TDC is that gain errors between TDCs in a multiple timing channel architecture will get progressively worse with increasing exposure times. Therefore – unless the TDCs can be locked to a stable oscillation frequency using a phase-locked loop (PLL) or similar – a mismatch compensation circuit is required at the output of each TDC, as well as a calibration procedure.

Although providing raw TCSPC codes at high data-rates would not allow efficient *real-time* calculation, it would act as a powerful tool for experimentation where accuracy is more important than processing speed (e.g. FLIM-FRET). The maximum bandwidth of a pad in this process is approximately 100 MHz, so assuming that 10-bit data is sufficient, each code can be transmitted off chip at up to 10 MHz through one pad. This defines the maximum excitation rate achievable before more than one pad is required per timing channel. An architecture to achieve raw TCSPC functionality is shown in Figure 6.4, where an on-chip PLL is locked to the incoming synchronisation pulse to provide the 100 MHz clock to a set of data serialisers. The PLL can also be used to lock the oscillation frequency of the TDCs to improve gain mismatch errors (V_{VCO}), whilst its output would be transmitted off-chip to act as the data sampling clock.

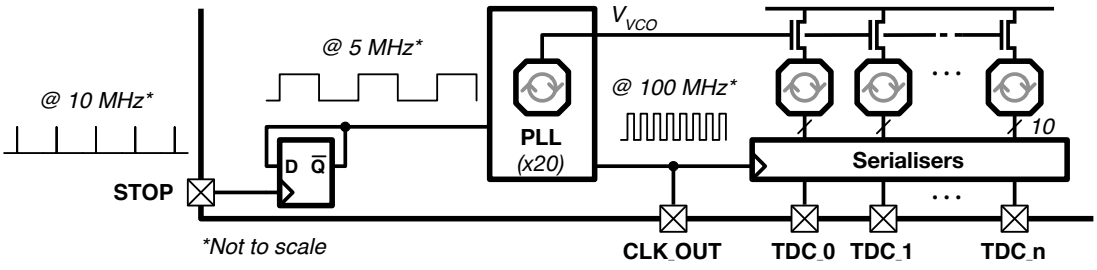


Figure 6.4: *Proposed multiple channel raw TCSPC architecture.*

6.5 Final Remarks

The initial motivations of this research were to increase the pile-up limited photon throughput of conventional single channel TCSPC apparatus and provide a means to extract fluorescence lifetime calculations in *real-time*, whilst miniaturising the multiple hardware components into a single device. These aims have been successfully realised by the design, implementation and demonstration of an advanced hardware/software system built around an integrated CMOS sensor core, that uses cutting edge SPAD, SiPM and TDC technology combined with *real-time* embedded CMM processing in a novel architecture. The evidence provided through modelling and simulation, as well as characterisation and experimental results successfully demonstrate that the sensor architecture is a significant advance in state of the art TCSPC and time-domain fluorescence lifetime instrumentation, providing photon throughputs in excess of the excitation frequency for minimal error in lifetime calculation by CMM.

It is hoped that the future direction of this work will branch into two primary themes: the development of experimental applications to exploit the performance gains and ease of use made possible by the technology; and continued investigation into an improved sensor architecture using the knowledge gained throughout this research. The most exciting future work promises to come from the integration of the device into experimental setups to advance cutting edge research in the fields of medical diagnosis and pharmacological development. This can be achieved not only for fluorescence lifetime based applications such as FLIM-FRET and flow based sorting; but also for other applications such as fNIRS and PET. Minor modifications to the sensor architecture and design – such as: adapting the SiPM to increase fill factor, improve the timing response and provide more output channels; using a *forward* mode timing and/or providing random TDC utilisation; and providing raw TCSPC data in addition the *real-time* CMM calculation – promise to provide even further performance improvements to this appealing new technology.

REFERENCES

[1] J. Arlt, D. Tyndall, B. R. Rae, D. D.-U. Li, J. A. Richardson, and R. K. Henderson, "A Study of Pile-up in Integrated Time-Correlated Single Photon Counting Systems," *Review of Scientific Instruments*, vol. 84, no. 10, Oct. 2013.

[2] D. Tyndall, B. R. Rae, D. D.-U. Li, J. Arlt, A. Johnston, J. A. Richardson, and R. K. Henderson, "A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13µm CMOS," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 6, pp. 562–570, Dec. 2012.

[3] D. Tyndall, B. R. Rae, D. D.-U. Li, J. A. Richardson, J. Arlt, and R. K. Henderson, "A 100Mphoton/s Time-Resolved Mini-Silicon Photomultiplier with On-Chip Fluorescence Lifetime Estimation in 0.13µm CMOS Imaging Technology," in *International Solid-State Circuits Conference (ISSCC)*, Feb. 2012, pp. 122–123.

[4] D. Tyndall, R. J. Walker, K. Nguyen, R. Galland, J. Gao, I. Wang, M. Kloster, A. Delon, and R. K. Henderson, "Automatic laser alignment for multifocal microscopy using a LCOS SLM and a 32×32 pixel CMOS SPAD array," in *Proceedings of SPIE*, vol. 8086, 2011.

[5] M. Kloster-Landsberg, D. Tyndall, I. Wang, R. J. Walker, R. K. Henderson, and A. Delon, "Note: Multi-confocal fluorescence correlation spectroscopy in living cells using a complementary metal oxide semiconductor-single photon avalanche diode array," *Review of Scientific Instruments*, vol. 84, no. 7, Jul. 2013.

[6] D. D.-U. Li, S. Poland, S. Coelho, D. Tyndall, W. Zhang, R. Walker, J. Richardson, and R. Henderson, "Advanced Fluorescence Lifetime Imaging Algorithms for CMOS Single-Photon Sensor Based Multi-focal Multi-photon Microscopy," in *Engineering in Medicine and Biology Conference (EMBC)*, Osaka, Jul. 2013.

[7] S. P. Poland, S. Coelho, N. Krstajic, D. Tyndall, R. J. Walker, J. Monypenny, D. D.-U. Li, R. K. Henderson, and S. Ameer-Beg, "Development of a Fast TCSPC FLIM-FRET Imaging System," in *Proceedings of SPIE*, vol. 8588, San Francisco, Feb. 2013.

[8] S. Coelho, S. P. Poland, N. Krstajic, D. D.-U. Li, J. Monypenny, R. J. Walker, D. Tyndall, T. C. Ng, R. K. Henderson, and S. Ameer-Beg, "Multibeam multiphoton microscopy

with adaptive optical correction," in *Proceedings of SPIE*, vol. 8588, San Francisco, Feb. 2013.

[9] D. D.-U. Li, S. Ameer-Beg, J. Arlt, D. Tyndall, R. J. Walker, D. R. Matthews, V. Visitkul, J. A. Richardson, and R. K. Henderson, "Time-Domain Fluorescence Lifetime Imaging Techniques Suitable for Solid-State Imaging Sensor Arrays," *Sensors*, vol. 12, no. 5, pp. 5650–5669, May 2012.

[10] D. D.-U. Li, J. Arlt, D. Tyndall, R. J. Walker, J. A. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm," *Journal of Biomedical Optics*, vol. 16, no. 9, p. 096012, Sep. 2011.

[11] G. Giraud, H. Schulze, D. D.-U. Li, T. T. Bachmann, J. Crain, D. Tyndall, J. A. Richardson, R. J. Walker, D. Stoppa, E. Charbon, R. K. Henderson, and J. Arlt, "Fluorescence lifetime biosensing with DNA microarrays and a CMOS-SPAD imager," *Biomedical Optics Express*, vol. 1, no. 5, pp. 1302–1308, Jan. 2010.

[12] J. A. Richardson, R. J. Walker, L. A. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, "A 32×32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," in *Custom Integrated Circuits Conference (CCIC)*, no. 029217, Sep. 2009, pp. 77–80.

[13] R. A. Colyer, G. Scalia, I. Rech, A. Gulinatti, M. Ghioni, S. Cova, S. Weiss, and X. Michalet, "High-throughput FCS using an LCOS spatial light modulator and an 8 × 1 SPAD array." *Biomedical Optics Express*, vol. 1, no. 5, pp. 1408–1431, Jan. 2010.

[14] J. R. Lakowicz, *Principals of Fluorescence Spectroscopy*, 3rd ed. Springer, 2006.

[15] R. Sanders, A. Draaijer, H. C. Gerritsen, P. M. Houpt, and Y. K. Levine, "Quantitative pH Imaging in Cells Using Confocal Fluorescence Lifetime Imaging Microscopy," *Analytical Biochemistry*, vol. 227, pp. 302–308, Sep. 1995.

[16] H. C. Gerritsen, R. Sanders, A. Draaijer, C. Ince, and Y. K. Levine, "Fluorescence lifetime imaging of oxygen in living cells," *Journal of Fluorescence*, vol. 7, no. 1, pp. 11–15, Mar. 1997.

[17] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, and M. L. Johnson, "Fluorescence lifetime imaging of calcium using Quin-2," *Cell Calcium*, vol. 13, no. 3, pp. 131–47, Mar. 1992.

[18] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*. Springer, 2005.

[19] D. V. O'Connor and D. Phillips, *Time-correlated Single Photon Counting*. Academic Press, 1984.

[20] C. M. Harris and B. K. Selinger, "Single-Photon Decay Spectroscopy II - The Pile-up Problem," *Austrailian Journal of Chemistry*, vol. 32, no. 10, pp. 2111–29, 1979.

[21] A. Esposito, "Beyond Range: Innovating Fluorescence Microscopy," *Remote Sensing*, vol. 4, no. 1, pp. 111–119, Jan. 2012.

[22] C. Veerappan, J. A. Richardson, R. J. Walker, D. D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter," in *International Solid-State Circuits Conference (ISSCC)*, Feb. 2011, pp. 312–313.

[23] A. Squire, P. J. Verveer, and P. I. Bastiaens, "Multiple frequency fluorescence lifetime imaging microscopy," *Journal of Microscopy*, vol. 197, no. 2, pp. 136–49, 2000.

[24] H. C. Gerritsen, M. A. H. Asselbergs, A. V. Agronskaia, and W. G. J. H. M. van Sark, "Fluorescence lifetime imaging in scanning microscopes: acquisition speed, photon economy and lifetime resolution." *Journal of Microscopy*, vol. 206, no. 3, pp. 218–24, Jun. 2002.

[25] R. M. Ballew and J. N. Demas, "An error analysis of the rapid lifetime determination method for the evaluation of single exponential decays," *Analytical Chemistry*, vol. 61, no. 1, pp. 30–33, 1989.

[26] M. Maus, M. Cotlet, J. Hofkens, T. Gensch, F. C. De Schryver, J. Schaffer, and C. a. Seidel, "An Experimental Comparison of the Maximum Likelihood Estimation and Nonlinear Least-Squares Fluorescence Lifetime Analysis of Single Molecules," *Analytical Chemistry*, vol. 73, no. 9, pp. 2078–86, May 2001.

[27] V. V. Apanasovich and E. G. Novikov, "Methods of Analysis of Fluorescence Decay Curves in Plused Fluorometry (Review)," *Journal of Applied Spectroscopy*, vol. 56, no. 4, pp. 317–327, 1992.

[28] A. T. N. Kumar, S. B. Raymond, B. J. Bacskai, and D. A. Boas, "Comparison of frequency-domain and time-domain fluorescence lifetime tomography," *Optics Letters*, vol. 33, no. 5, pp. 470–2, Mar. 2008.

[29] E. Gratton, S. Breusegem, J. Sutin, Q. Ruan, and N. Barry, "Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods." *Journal of Biomedical Optics*, vol. 8, no. 3, pp. 381–90, Jul. 2003.

[30] A. Esposito, H. C. Gerritsen, and F. S. Wouters, "Fluorescence lifetime heterogeneity resolution in the frequency domain by lifetime moments analysis." *Biophysical Journal*, vol. 89, no. 6, pp. 4286–99, Dec. 2005.

[31] ——, "Optimizing frequency-domain fluorescence lifetime sensing for high-throughput applications: photon economy and acquisition speed," *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, vol. 24, no. 10, pp. 3261–73, Oct. 2007.

[32] R. A. Colyer, C. Lee, and E. Gratton, "A Novel Fluorescence Lifetime Imaging System That Optimizes Photon Efficiency," *Microscopy Research and Technique*, vol. 71, no. 3, pp. 201–213, 2008.

[33] D. Mosconi, D. Stoppa, L. Pancheri, L. Gonzo, and A. Simoni, "CMOS Single-Photon Avalanche Diode Array for Time-Resolved Fluorescence Detection," in *European Solid State Circuits Conference (ESSCIRC)*, Sep. 2006, pp. 564–567.

[34] B. R. Rae, J. Yang, J. J. D. McKendry, Z. Gong, D. Renshaw, J. M. Girkin, E. Gu, M. D. Dawson, and R. K. Henderson, "A Vertically Integrated CMOS Microsystem for Time-Resolved Fluorescence Analysis," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 4, no. 6, pp. 437–444, Dec. 2010.

[35] L. Pancheri and D. Stoppa, "A SPAD-based Pixel Linear Array for High-Speed Time-Gated Fluorescence Lifetime Imaging," in *European Solid State Circuits Conference (ESSCIRC)*, Sep. 2009, pp. 428–431.

[36] D. S. Elson, S. Webb, J. Siegel, K. Suhling, D. Davis, M. J. Lever, D. Phillips, A. Wallace, and P. M. W. French, "Biomedical Applications of Fluorescence Lifetime Imaging," *Optics and Photonics News*, pp. 26–57, 2002.

[37] A. G. Ryder, S. Power, T. J. Glynn, and J. J. Morrison, "Time-domain measurement of fluorescence lifetime variation with pH," in *Proceedings of SPIE*, vol. 4259, 2001.

[38] W. Zhong, P. Urayama, and M.-A. Mycek, "Imaging fluorescence lifetime modulation of a ruthenium-based dye in living cells: the potential for oxygen sensing," *Journal of Physics D: Applied Physics*, vol. 36, no. 14, pp. 1689–1695, Jul. 2003.

[39] A. Takahashi, P. Camacho, J. D. Lechleiter, and B. Herman, "Measurement of intracellular Calcium," *Physiological Reviews*, vol. 79, no. 4, pp. 1089–125, Oct. 1999.

[40] Q. Zhao, I. T. Young, and J. G. S. de Jong, "Photon budget analysis for fluorescence lifetime imaging microscopy," *Journal of Biomedical Optics*, vol. 16, no. 8, p. 086007, Aug. 2011.

[41] D. M. Shotton, "Confocal scanning optical microscopy and its applications for biological specimens," *Journal of Cell Science*, vol. 94, pp. 175–206, 1989.

[42] F. Helmchen and W. Denk, "Deep tissue two-photon microscopy," *Nature Methods*, vol. 2, no. 12, pp. 932–41, 2005.

[43] H.-J. Yoon, S. Itoh, and S. Kawahito, "A CMOS Image Sensor With In-Pixel Two-Stage Charge Transfer for Fluorescence Lifetime Imaging," *IEEE Transactions on Electron Devices*, vol. 56, no. 2, pp. 214–221, 2009.

[44] A. Esposito, H. C. Gerritsen, T. Oggier, F. Lustenberger, and F. S. Wouters, "Innovating lifetime microscopy: a compact and simple tool for life sciences, screening, and diagnostics." *Journal of Biomedical Optics*, vol. 11, no. 3, p. 34016, 2006.

[45] R. R. Duncan, A. Bergmann, M. A. Cousin, D. K. Apps, and M. J. Shipston, "Multi-Dimensional Time-Correlated Single Photon Counting (TCSPC) Fluorescence Lifetime Imaging Microscopy (FLIM) to Detect FRET in Cells," *Journal of Microscopy*, vol. 215, no. 1, pp. 1–12, 2004.

[46] H. Wallrabe and A. Periasamy, "Imaging protein molecules using FRET and FLIM microscopy." *Current Opinion in Biotechnology*, vol. 16, no. 1, pp. 19–27, Feb. 2005.

[47] S. Turconi, R. P. Bingham, U. Haupts, and A. J. Pope, "Developments in fluorescence lifetime-based analysis for ultra-HTS," *Drug Discovery Today*, vol. 6, no. 12, pp. 27–39, 2001.

[48] A. Esposito, C. P. Dohm, M. Bähr, and F. S. Wouters, "Unsupervised fluorescence lifetime imaging microscopy for high content and high throughput screening." *Molecular and Cellular Proteomics*, vol. 6, no. 8, pp. 1446–54, Aug. 2007.

[49] C. D'Andrea, A. Bassi, P. Taroni, D. Pezzoli, A. Volonterio, and G. Candiani, "Time-resolved fluorescence spectroscopic investigation of cationic polymer/DNA complex formation," in *Proceedings of SPIE*, vol. 8087, 2011.

[50] P. R. Barber, S. Ameer-Beg, J. Gilbey, R. J. Edens, I. Ezike, and B. Vojnovic, "Global and pixel kinetic data analysis for FRET detection by multi-photon time-domain FLIM," in *Proceedings of SPIE*, vol. 5700, 2005.

[51] M. J. Fulwyler, "Status Quo in Flow-Through Cytometry," *Journal of Histochemistry and Cytochemistry*, vol. 22, no. 7, pp. 605–606, Jul. 1974.

[52] W. A. Bonner, H. R. Hulett, R. G. Sweet, and L. A. Herzenberg, "Fluorescence Activated Cell Sorting," *Review of Scientific Instruments*, vol. 43, no. 3, pp. 404–409, 1972.

[53] J. P. Houston, M. A. Naivar, and J. P. Freyer, "Digital Analysis and Sorting of Fluorescence Lifetime by Flow Cytometry," *Cytometry Part A*, vol. 77, no. 9, pp. 861–872, Sep. 2010.

[54] M. Wahl, R. Erdmann, K. Lauritsen, and H.-J. Rahn, "Hardware solution for continuous time-resolved burst detection of single molecules in flow," in *Proceedings of SPIE*, Apr. 1998, pp. 173–178.

[55] J. F. Keij and J. A. Steinkamp, "Flow cytometric characterization and classification of multiple dual-color fluorescent microspheres using fluorescence lifetime." *Cytometry*, vol. 33, no. 3, pp. 318–23, Nov. 1998.

[56] H. H. Cui, J. G. Valdez, J. A. Steinkamp, and H. A. Crissman, "Fluorescence lifetime-based discrimination and quantification of cellular DNA and RNA with phase-sensitive flow cytometry." *Cytometry Part A*, vol. 52, no. 1, pp. 46–55, Mar. 2003.

[57] M. Hintersteiner, T. Kimmerlin, F. Kalthoff, M. Stoeckli, G. Garavel, J.-M. Seifert, N.-C. Meisner, V. Uhl, C. Buehler, T. Weidemann, and M. Auer, "Single bead labeling method for combining confocal fluorescence on-bead screening and solution validation of tagged one-bead one-compound libraries." *Chemistry and Biology*, vol. 16, no. 7, pp. 724–35, Jul. 2009.

[58] C. D. Salthouse, R. Weissleder, and U. Mahmood, "Development of a Time Domain Fluorimeter for Fluorescent Lifetime Multiplexing Analysis," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 2, no. 3, pp. 204–211, Sep. 2008.

[59] J. A. Richardson, L. A. Grant, and R. K. Henderson, "Low Dark Count Single-Photon Avalanche Diode Structure Compatible With Standard Nanometer Scale CMOS Technology," *IEEE Photonics Technology Letters*, vol. 21, no. 14, pp. 1020–1022, Jul. 2009.

[60] J. Yguerabide, "Nanosecond Fluorescence Spectroscopy of Macromolecules," *Methods in Enzymology*, vol. 26, pp. 498–578, 1972.

[61] A. E. W. Knight and B. K. Selinger, "Single Photon Decay Spectroscopy," *Austrailian Journal of Chemistry*, vol. 26, no. 1, pp. 1–27, 1973.

[62] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, "Invited review article: Single-photon sources and detectors." *Review of Scientific Instruments*, vol. 82, no. 7, Jul. 2011.

[63] G. Hungerford and D. J. S. Birch, "Single-photon timing detectors for fluorescence lifetime spectroscopy," *Measurement Science and Technology*, vol. 7, no. 2, pp. 121–35, 1996.

[64] D. J. S. Birch, "Fluorescence detections and directions," *Measurement Science and Technology*, vol. 22, no. 5, May 2011.

[65] U. Akgun, A. S. Ayan, G. Aydin, F. Duru, J. Olson, and Y. Onel, "Afterpulse timing and rate investigation of three different Hamamatsu Photomultiplier Tubes," *Journal of Instrumentation*, vol. 3, no. 01, Jan. 2008.

[66] J. Kalisz, "Review of methods for time interval measurements with picosecond resolution," *Metrologia*, vol. 41, no. 1, pp. 17–32, Feb. 2004.

[67] G. W. Roberts and M. Ali-Bakhshian, "A Brief Introduction to Time-to-Digital and Digital-to-Time Converters," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 57, no. 3, pp. 153–157, Mar. 2010.

[68] F. Koberling, B. Kramer, S. Tannert, S. Rüttinger, U. Ortmann, M. Patting, M. Wahl, B. Ewers, P. Kapusta, and R. Erdmann, "Recent advances in time-correlated single-photon counting," in *Proceedings of SPIE*, vol. 6862, 2008.

[69] M. Patting, M. Wahl, P. Kapusta, and R. Erdmann, "Dead-time effects in TCSPC data analysis," in *Proceedings of SPIE*, vol. 6583, 2007.

[70] P. B. Coates, "The correction for photon 'pile-up' in the measurement of radiative lifetimes," *Journal of Physics E: Scientific Instruments*, vol. 878, pp. 4–6, 1968.

[71] J. G. Walker, "Iterative correction for 'pile-up' in single-photon lifetime measurement," *Optics Communications*, vol. 201, pp. 271–277, 2002.

[72] C. C. Davis and T. A. King, "Correction methods for photon pile-up in lifetime determination by single-photon counting," *Journal of Physics A: Mathematical and Theoretical*, vol. 3, pp. 101–9, 1970.

[73] O. M. Williams and W. J. Sandle, "A 'pile-up' gate generator for removing distortion in multichannel delayed coincidence experiments," *Journal of Physics E: Scientific Instruments*, vol. 3, no. 9, pp. 741–743, Sep. 1970.

[74] M. Wahl, H.-J. Rahn, I. Gregor, R. Erdmann, and J. Enderlein, "Dead-time optimized time-correlated photon counting instrument with synchronized, independent timing channels." *Review of Scientific Instruments*, vol. 78, no. 3, p. 033106, Mar. 2007.

[75] D. McLoskey, D. Campbell, A. Allison, and G. Hungerford, "Fast time-correlated single-photon counting fluorescence lifetime acquisition using a 100 MHz semiconductor excitation source," *Measurement Science and Technology*, vol. 22, no. 6, p. 067001, Apr. 2011.

[76] W. Becker, A. Bergmann, G. Biscotti, and C. Biskup, "Fluorescence Lifetime Imaging by Multi-Detector TCSPC," in *Biomedical Topical Meeting (BIO)*, Miami Beach, Florida, Apr. 2004.

[77] W. Becker, A. Bergmann, E. Haustein, Z. Petrasek, P. Schwille, C. Biskup, L. Kelbauskas, K. Benndorf, N. Klöcker, T. Anhut, I. Riemann, and K. König, "Fluorescence Lifetime Images and Correlation Spectra Obtained by Multidimensional Time-Correlated Single Photon Counting," *Microscopy Research and Technique*, vol. 69, no. 3, pp. 186–95, Mar. 2006.

[78] D. E. Schwartz, E. Charbon, and K. L. Shepard, "A Single-Photon Avalanche Diode Array for Fluorescence Lifetime Imaging Microscopy," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2546–2557, 2008.

[79] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 $\times$ 128 Single-Photon Image Sensor With Column-Level 10-Bit Time-to-Digital Converter Array," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, 2008.

[80] W. Becker, A. Bergmann, H. Wabnitz, D. Grosenick, and A. Liebert, "High count rate multichannel TCSPC for optical tomography," in *Biomedical Topical Meeting (BIO)*, Miami Beach, Florida, Apr. 2002.

[81] M. Wahl, H.-J. Rahn, T. Röhlicke, G. Kell, D. Nettels, F. Hillger, B. Schuler, and R. Erdmann, "Scalable time-correlated photon counting system with multiple independent input channels," *Review of Scientific Instruments*, vol. 79, no. 12, p. 123113, Dec. 2008.

[82] M. Wahl, G. Kell, P. Kapusta, H.-J. Rahn, T. Roehlicke, and R. Erdmann, "New multichannel photon timing instrumentation with independent synchronized channels and high count rate for FLIM and correlation analysis," in *Proceedings of SPIE*, vol. 7183, 2009.

[83] S. Donati, G. Martini, and E. Randone, "Improving Photodetector Performance by Means of Microoptics Concentrators," *Journal of Lightwave Technology*, vol. 29, no. 5, pp. 661–665, Mar. 2011.

[84] M. Gösch, H. Blom, S. Anderegg, K. Korn, P. Thyberg, M. Wells, T. Lasser, R. Rigler, A. Magnusson, and S. Hård, "Parallel dual-color fluorescence cross-correlation spectroscopy using diffractive optical elements." *Journal of biomedical optics*, vol. 10, no. 5, 2005.

[85] A. Rochas, "Single Photon Avalanche Diodes in CMOS Technology," Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), 2003.

[86] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Applied Optics*, vol. 35, no. 12, pp. 1956–76, Apr. 1996.

[87] S. Pellegrini, R. Warburton, L. Tan, J. Ng, A. Krysa, K. Groom, J. David, S. Cova, M. Robertson, and G. Buller, "Design and Performance of an InGaAs-InP Single-Photon Avalanche Diode Detector," *IEEE Journal of Quantum Electronics*, vol. 42, no. 4, pp. 397–403, Apr. 2006.

[88] K. A. McIntosh, J. P. Donnelly, D. C. Oakley, A. Napoleone, S. D. Calawa, L. J. Mahoney, K. M. Molvar, J. Mahan, R. J. Molnar, E. K. Duerr, G. W. Turner, M. J. Manfra, and B. F. Aull, "Arrays of III-V semiconductor Geiger-mode avalanche photodiodes," in *Lasers and Electro-Optics Society (LEOS)*, vol. 2, 2003, pp. 686–7.

[89] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance," *Electronics Letters*, vol. 25, no. 13, pp. 841–3, 1989.

[90] X. Llopart, M. Campbell, D. San Segundo, E. Pernigotti, and R. Dinapoli, "Medipix2, a 64k pixel read out chip with 55 μm square elements working in single photon counting mode," in *Nuclear Science Symposium Conference Record (NSS/MIC)*, vol. 3, 2001, pp. 1484–1488.

[91] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 863–869, 2007.

[92] M. Gersbach, C. Niclass, E. Charbon, J. A. Richardson, R. K. Henderson, and L. A. Grant, "A Single Photon Detector Implemented in a 130nm CMOS Imaging Process," in *European Solid State Device Research Conference (ESSDERC)*, no. 029217, 2008, pp. 270–273.

[93] H. Finkelstein, M. J. Hsu, and S. C. Esener, "STI-Bounded Single-Photon Avalanche Diode in a Deep-Submicrometer CMOS Technology," *IEEE Electron Device Letters*, vol. 27, no. 11, pp. 887–9, 2006.

[94] N. Faramarzpour, M. J. Deen, S. Shirani, and Q. Fang, "Fully Integrated Single Photon Avalanche Diode Detector in Standard CMOS 0.18- μm Technology," *IEEE Transactions on Electron Devices*, vol. 55, no. 3, pp. 760–7, 2008.

[95] M. A. Marwick and A. G. Andreou, "Single photon avalanche photodetector with integrated quenching fabricated in TSMC 0.18 μm 1.8 V CMOS process," *Electronics Letters*, vol. 44, no. 10, pp. 643–4, 2008.

[96] D. D.-U. Li, J. Arlt, J. A. Richardson, R. J. Walker, A. Buts, D. Stoppa, E. Charbon, and R. K. Henderson, "Real-time fluorescence lifetime imaging system with a 32×32 0.13μm CMOS low dark-count single-photon avalanche diode array," *Optics Express*, vol. 18, no. 10, pp. 10 257–69, 2010.

[97] M. Cohen, F. Roy, D. Herault, Y. Cazaux, A. Gandolfi, J. P. Reynard, C. Cowache, E. Bruno, T. Girault, J. Vaillant, F. Barbier, Y. Sanchez, N. Hotellier, O. Leborgne, C. Augier, A. Inard, T. Jagueneau, J. Michailos, and E. Mazaleyrat, "Fully Optimized Cu based process with dedicated cavity etch for 1.75µm and 1.45µm pixel pitch CMOS Image Sensors," in *International Electron Devices Meeting (IEDM)*, Dec. 2006.

[98] F. Zappa, M. Ghioni, S. Cova, C. Samori, and A. C. Giudice, "An Integrated Active-Quenching Circuit for Single-Photon Avalanche Diodes," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 6, pp. 1167–1175, 2000.

[99] A. Eisele, R. K. Henderson, B. Schmidtke, T. Funk, L. A. Grant, J. A. Richardson, and W. Freude, "185 MHz Count Rate, 139 dB Dynamic Range Single-Photon Avalanche Diode with Active Quenching Circuit in 130 nm CMOS Technology," in *International Image Sensor Workshop (IISW)*, 2011.

[100] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable Single-Photon Avalanche Diode Structures in Nanometer CMOS Technology," *IEEE Transactions on Electron Devices*, vol. 58, no. 7, pp. 2028–2035, Jul. 2011.

[101] F. Zappa, A. Gulinatti, P. Maccagnani, S. Tisa, and S. Cova, "SPADA: Single-Photon Avalanche Diode Arrays," *IEEE Photonics Technology Letters*, vol. 17, no. 3, pp. 657–659, Mar. 2005.

[102] K. Yamamoto, K. Yamamura, K. Sato, T. Ota, H. Suzuki, and S. Ohsuka, "Development of Multi-Pixel Photon Counter (MPPC)," *Nuclear Science Symposium Conference Record (NSS/MIC)*, pp. 1094–1097, 2006.

[103] M. Mazzillo, G. Condorelli, D. Sanfilippo, G. Valvo, B. Carbone, G. Fallica, S. Billotta, M. Belluso, G. Bonanno, L. Cosentino, A. Pappalardo, and P. Finocchiaro, "Silicon Photomultiplier Technology at STMicroelectronics," *IEEE Transactions on Nuclear Science*, vol. 56, no. 4, pp. 2434–2442, Aug. 2009.

[104] N. Otte, "The Silicon Photomultiplier - A new device for High Energy Physics, Astroparticle Physics, Industrial and Medical Applications," in *International Symposium on Detector Development for Particle, Astroparticle and Synchrotron Radiation Experiments (SNIC)*, no. April, 2006.

[105] J. F. Christian, C. J. Stapels, E. B. Johnson, M. McClish, P. Dokhale, K. S. Shah, S. Mukhopadhyay, E. Chapman, and F. L. Augustine, "Advances in CMOS solid-state photomultipliers for scintillation detector applications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 624, no. 2, pp. 449–458, Dec. 2010.

[106] E. Grigoriev, A. Akindinov, M. Breitenmoser, S. Buono, E. Charbon, C. Niclass, I. Desforges, and R. Rocca, "Silicon photomultipliers and their bio-medical applications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 571, no. 1-2, pp. 130–133, Feb. 2007.

[107] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier — Principle of operation and intrinsic detector performance," in *Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct. 2009, pp. 1959–1965.

[108] S. Mandai, V. Jain, and E. Charbon, "A Fully-Integrated 780×800μm^2 Multi-Digital Silicon Photomultiplier With Column-parallel Time-to-Digital Converter," in *European Solid State Circuits Conference (ESSCIRC)*, 2012.

[109] G. Prescher and T. Frach, "Photodiode Self-Test," *US Patent 8,193,815 B2*, 2012.

[110] J. Heller and J. Breisch, "CMOS Imaging Device with Integrated Defective Pixel Correction Circuitry," *US Patent 6,396,539 B1*, 2002.

[111] K. Dong, "On-Chip Dead Pixel Correction in a CMOS Imaging Sensor," *US Patent 6,665,009 B1*, 2003.

[112] R. J. Walker, E. A. G. Webster, J. Li, N. Massari, and Robert K. Henderson, "High Fill Factor Digital Silicon Photomultiplier Structures in 130nm CMOS Imaging Technology," in *Nuclear Science Symposium Conference Record (NSS/MIC)*, Nov. 2012.

[113] E. A. G. Webster, R. J. Walker, Robert K. Henderson, and L. A. Grant, "A silicon photomultiplier with >30% detection efficiency from 450-750nm and 11.6μm pitch NMOS-only pixel with 21.6% fill factor in 130nm CMOS," in *European Solid State Device Research Conference (ESSDERC)*, 2012.

[114] L. H. C. Braga, L. Pancheri, L. Gasparini, M. Perenzoni, R. J. Walker, R. K. Henderson, and D. Stoppa, "A CMOS mini-SiPM detector with in-pixel data compression for PET applications," in *Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct. 2011, pp. 548–552.

[115] J. A. Richardson, "Time Resolved Single Photon Imaging in Nanometer Scale CMOS Technology," Ph.D. dissertation, The University of Edinburgh, 2010.

[116] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps Time-to-Digital Converter for Frequency Synthesis in 90-nm CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 3, pp. 220–224, Mar. 2006.

[117] P. Dudek, S. Szczepanski, and J. V. Hatfield, "A High-Resolution CMOS Time-to-Digital Converter Utilizing a Vernier Delay Line," *IEEE Transactions on Solid-State Circuits*, vol. 35, no. 2, pp. 240–247, 2000.

[118] K. Karadamoglou, N. P. Paschalidis, E. Sarris, N. Stamatopoulos, G. Kottaras, and V. Paschalidis, "An 11-bit High-Resolution and Adjustable-Range CMOS Time-to-Digital Converter for Space Science Instruments," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 1, pp. 214–222, Jan. 2004.

[119] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-landsiedel, "A Local Passive Time Interpolation Concept for Variation-Tolerant High-Resolution Time-to-Digital Conversion," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, 2008.

[120] M. Lee and A. A. Abidi, "A 9b, 1.25ps Resolution Coarse-Fine Time-to-Digital Converter in 90nm CMOS that Amplifies a Time Residue," in *Symposium on VLSI Circuits*, 2007, pp. 68–69.

[121] R. Nutt, "Digital Time Intervalometer," *Review of Scientific Instruments*, vol. 39, no. 9, p. 1342, 1968.

[122] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa, J. A. Richardson, R. J. Walker, R. K. Henderson, and E. Charbon, "A Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated in Deep-Submicron CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 6, pp. 1394–1407, 2012.

[123] B. K. Swann, B. J. Blalock, L. G. Clonts, D. M. Binkley, J. M. Rochelle, E. Breeding, and K. M. Baldwin, "A 100-ps Time-Resolution CMOS Time-to-Digital Converter for Positron Emission Tomography Imaging Applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 11, pp. 1839–1852, 2004.

[124] I. Nissinen, A. Mäntyniemi, and J. Kostamovaara, "A CMOS Time-to-Digital Converter based on a Ring Oscillator for a Laser Radar," in *European Solid State Circuits Conference (ESSCIRC)*, vol. 29, 2003.

[125] B. M. Helal, M. Z. Straayer, G.-Y. Wei, and M. H. Perrott, "A Low Jitter 1.6 GHz Multiplying DLL Utilizing a Scrambling Time-to-Digital Converter and Digital Correlation," in *IEEE Symposium on VLSI Circuits*. Ieee, Jun. 2007, pp. 166–167.

[126] C. Vogel and H. Johansson, "Time-Interleaved Analog-To-Digital Converters: Status and Future Directions," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006, pp. 3386–3389.

[127] S. Danesh, J. Hurwitz, K. Findlater, D. Renshaw, and R. K. Henderson, "A Reconfigurable 1 GSps to 250 MSps , 7-bit to 9-bit Highly Time-Interleaved Counter ADC with Low Power Comparator Design," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 1–16, 2013.

[128] C. Vogel, M. Hotz, S. Saleem, K. Hausmair, and M. Soudan, "A Review on Low-Complexity Structures and Algorithms for the Correction of Mismatch Errors in Time-interleaved ADCs," in *IEEE International NEWCAS Conference*, 2012.

[129] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[130] P. Hall and B. Selinger, "Better estimates of exponential decay parameters," *The Journal of Physical Chemistry*, vol. 85, no. 20, pp. 2941–2946, Oct. 1981.

[131] C. J. de Grauw and H. C. Gerritsen, "Multiple Time-Gate Module for Fluorescence Lifetime Imaging," *Applied Spectroscopy*, vol. 55, no. 6, pp. 670–678, Jun. 2001.

[132] D. D.-U. Li, E. Bonnist, D. Renshaw, and R. K. Henderson, "On-chip, time-correlated, fluorescence lifetime extraction algorithms and error analysis," *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, vol. 25, no. 5, pp. 1190–1198, 2008.

[133] D. D.-U. Li, R. J. Walker, J. A. Richardson, B. R. Rae, A. Buts, D. Renshaw, and R. K. Henderson, "FPGA Implementation of a Video-rate Fluorescence Lifetime Imaging System with a 32×32 CMOS Single-Photon Avalanche Diode Array," in *International Symposium on Circuits and Systems (ISCAS)*, vol. 1, 2009.

[134] D. D.-U. Li, B. R. Rae, R. Andrews, J. Arlt, and R. K. Henderson, "Hardware implementation algorithm and error analysis of high-speed fluorescence lifetime sensing systems using center-of-mass method," *Journal of Biomedical Optics*, vol. 15, no. 1, p. 017006, Jan. 2010.

[135] D. S. Elson, I. Munro, J. Requejo-Isidro, J. McGinty, C. Dunsby, N. Galletly, G. W. H. Stamp, M. A. A. Neil, M. J. Lever, P. A. Kellett, A. Dymoke-Bradshaw, J. Hares, and P. M. W. French, "Real-time time-domain fluorescence lifetime imaging including single-shot acquisition with a segmented optical image intensifier," *New Journal of Physics*, vol. 6, p. 180, Nov. 2004.

[136] D. S. Elson, J. Siegel, S. E. D. Webb, S. Lévêque-Fort, M. J. Lever, P. M. W. French, K. Lauritsen, M. Wahl, and R. Erdmann, "Fluorescence lifetime system for microscopy and multiwell plate imaging with a blue picosecond diode laser." *Optics Letters*, vol. 27, no. 16, pp. 1409–11, Aug. 2002.

[137] J. Doernberg, H.-S. Lee, and D. A. Hodges, "Full-Speed Testing of A/D Converters," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 6, pp. 820–827, 1984.

[138] A. Leff and J. T. Rayfield, "Web-Application Development Using the Model/View/Controller Design Pattern," in *Enterprise Distributed Object Computing (EDOC) Conference*, 2001, pp. 118–127.

[139] L. Pancheri and D. Stoppa, "A low-cost picosecond laser module for time- resolved optical sensing applications," *IEEE Sensors Journal*, vol. 11, no. 6, pp. 1380–1381, Jun. 2010.

[140] Y. Wang, B. R. Rae, R. K. Henderson, Z. Gong, J. J. D. McKendry, E. Gu, M. D. Dawson, G. A. Turnbull, and I. D. W. Samuel, "Ultra-portable explosives sensor based on a CMOS fluorescence lifetime analysis micro-system," *AIP Advances*, vol. 1, p. 032115, 2011.

[141] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, "An Infra-Red Sensitive, Low Noise, Single-Photon Avalanche Diode in 90nm CMOS," in *International Image Sensor Workshop (IISW)*, 2011.

[142] A. McCarthy, R. J. Collins, N. J. Krichel, V. Fernández, A. M. Wallace, and G. S. Buller, "Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting." *Applied optics*, vol. 48, no. 32, pp. 6241–51, Nov. 2009.

[143] D. Contini, A. Torricelli, A. Pifferi, L. Spinelli, F. Paglia, and R. Cubeddu, "Multi-channel time-resolved system for functional near infrared spectroscopy." *Optics Express*, vol. 14, no. 12, pp. 5418–32, Jun. 2006.

[144] L. H. C. Braga, L. Gasparini, L. A. Grant, R. K. Henderson, N. Massari, M. Perenzoni, and R. J. Walker, "An 8×16-pixel 92kSPAD Time-Resolved Sensor with On-Pixel 64ps 12b TDC and 100MS/s Real-Time Energy Histogramming in 0.13µm CIS Technology for PET/MRI Applications," in *International Solid-State Circuits Conference (ISSCC)*, San Francisco, Feb. 2013.

[145] G. O. Fruhwirth, S. Ameer-Beg, R. Cook, T. Watson, T. Ng, and F. Festy, "Fluorescence lifetime endoscopy using TCSPC for the measurement of FRET in live cells." *Optics Express*, vol. 18, no. 11, pp. 11 148–58, May 2010.

A
APPENDICES

A.1 SPAD Characteristics

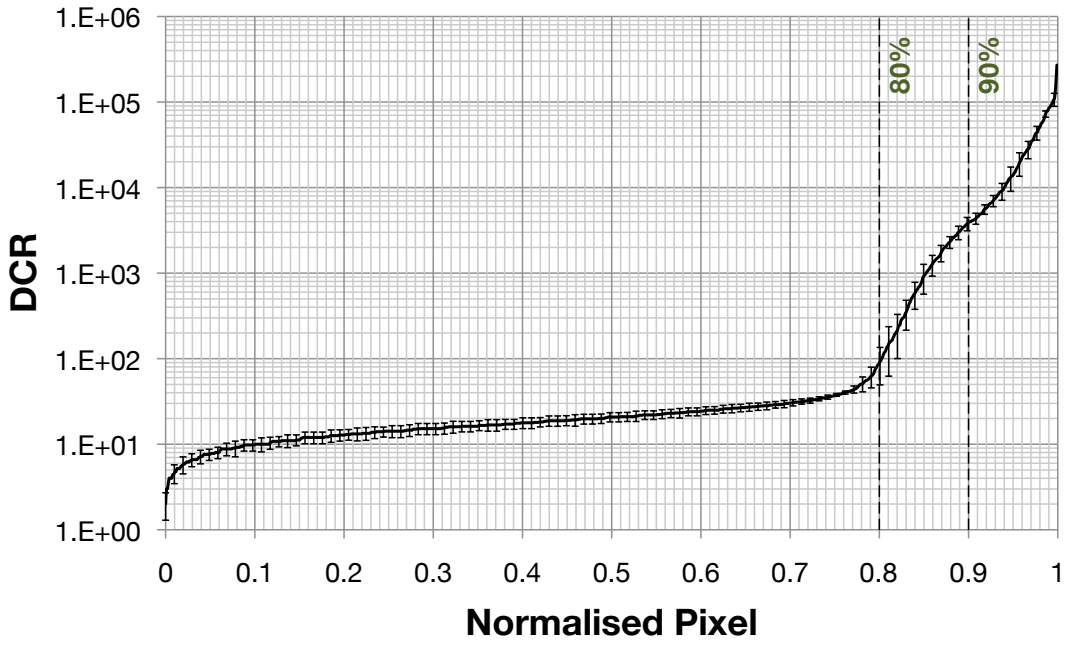


Figure A.1: *Ordered DCR distribution of multiple SPAD devices from [12].*

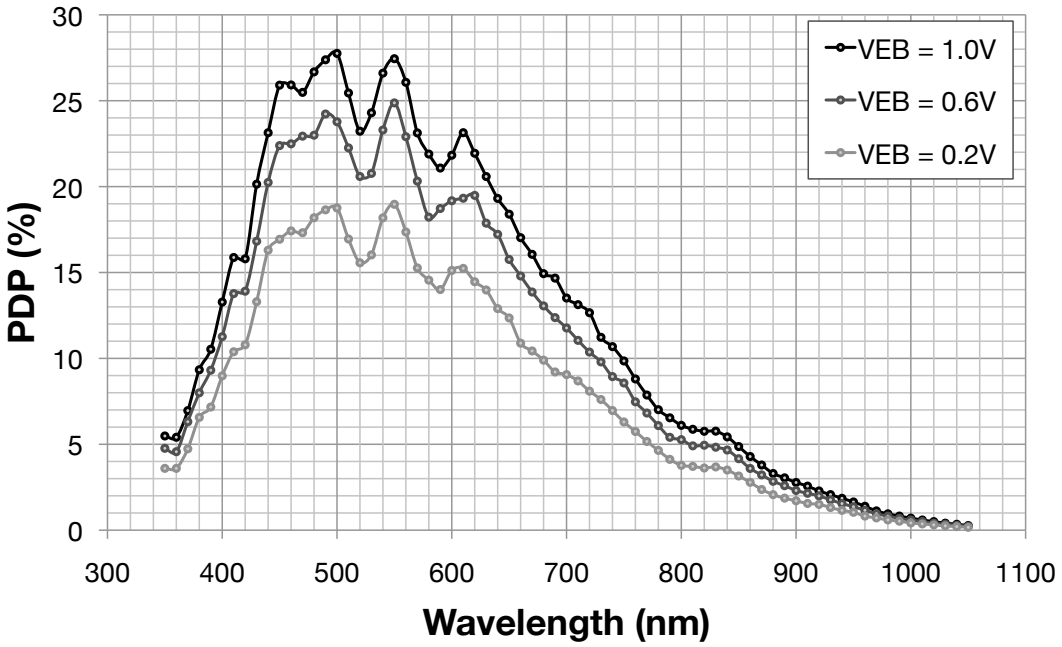


Figure A.2: *Photon detection probability (PDP) of the SPAD from [12].*

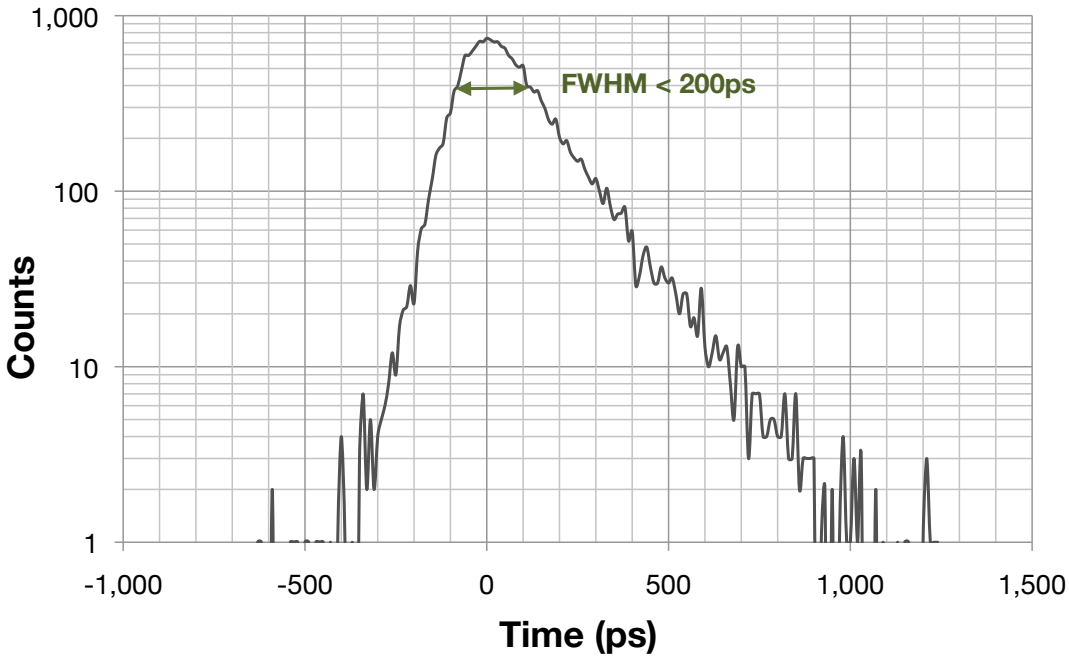


Figure A.3: *Timing jitter of the SPAD from [12].*

A.2 Pile-Up Model

```
%% Function Declaration
function [t2,...
          decay2,...
          decay2_PU,...
          cmm,...
          cmm_PU,...
          decay2_PU_ch,...
          PU_D,...
          PU_T,...
          PU_P]...
  = PileUpModel(C_P,...
                tau,...
                mu,...
                DCR,...
                t_P,...
                N_T,...
                N_D,...
                t_D,...
                sig_T,...
                N_M,...
                rev,...
                T_P)


%% Fixed Variables

% Laser repitition rate
f_E = 20.0*1000000;

% Laser period (ps)
per = (1e12/f_E);

% TDC R_T (ps)
R_T = 50.0;
```

```
41 %% Create Ideal lifetime decay
42
43 % Approximate peak counts in ideal decay (@ 1ps)
44 C_P2 = C_P/R_T;
45
46 % Time (@ 1 ps)
47 t1 = 0:per-1;
48
49 % Ideal histogram
50 decay1 = C_P2*(1./exp(t1/tau));
51
52
53 %% Add shot-noise and DCR (uncorrelated light)
54
55 % Discrete Ideal histogram + shot-noise
56 decay1 = round( poissrnd(decay1) );
57
58 % Total number of photon events
59 N_events = sum(decay1);
60
61 % Number of excitation periods (macro times)
62 N_per = N_events/mu;
63
64 % Total time of experiment (in seconds)
65 time = N_per / f_E;
66
67 % Create poisson distributed white-noise
68 noise = poissrnd( (DCR*time)/per, 1, per );
69
70 % Add histogram and noise floor
71 decay1 = decay1 + noise;
72
73 % Re-calculate total number of photon events
74 N_events = sum(decay1);
75
76
77
78
79
80
81
```

```
82 %% Convert histogram decay data to individual discrete events
83
84 % Create array to hold events
85 events = zeros(1,N_events);
86
87 % Index to count through total number of events
88 index = 1;
89
90 % Loop through each bin in histogram
91 for i=1:per
92   % Loop through each event within current bin
93   for j=1:decay1(1,i)
94     % If reverse start-stop
95     if rev == 1
96       % Add event at current index
97       events(index) = mod( (per - i) + T_P, per );
98     else
99       % Add event at current index
100      events(index) = mod( i + T_P, per );
101    end
102    % Increment Index
103    index = index + 1;
104  end
105 end
106
107
108 %% Assign events to excitation period (temporal)
109
110 % Create array of random excitation period indexes
111 pos = ceil( (N_per-1)*rand(1,N_events) );
112
113 % Sort random positions
114 % [ sorted array, index of sorted items] = sort( X )
115 [ sorted order ] = sort( pos );
116
117 % Order by positions
118 % NB. Order within each time window will remain
119 events(1,:) = events(1,order);
120
121 % Add sorted positions to events array
122 events = [ events; sorted ];
```

```
123 %% Assign events to SPAD detectors (spatial)
124
125 % Add random SPAD id to events array
126 events = [ events; ceil( N_D*rand(1,N_events) ) ];
127
128 % Order by same positions as above
129 events(3,:) = events(3,order);
130
131 clear pos order sorted;
132
133
134 %% Process events assuming no pile-up
135
136 % Create x-axis for processed decay histogram
137 t2 = (R_T:R_T:per+R_T);
138
139 % Calculate the number of bins in the histogram
140 bins = size(t2,2);
141
142 % Create array to hold histogram for ideal decay
143 decay2 = zeros(1,bins);
144
145 % Loop around all events
146 for i=1:N_events
147   % Calculate the histogram bin to be incrememented
148   bin = ceil( events(1,i) / R_T );
149   % Increment location in decay indexed by bin
150   decay2(1, bin) = decay2(1, bin) + 1;
151 end
152
153
154
155
156
157
158
159
160
161
162
163
```

```
164 %% Process events assuming pile-up
165
166 % If reverse start-stop, set negate flag
167 if (rev == 1)
168   neg = -1;
169 else
170   neg = 1;
171 end
172
173 % Create array of TDC resolutions using variations
174 R_T2 = R_T + sig_T;
175
176 % Create arrays to store decay data for each timing channel
177 decay2_PU_ch = zeros(N_T*N_M,bins);
178
179 % Detector, SiPM pulse-width and timing-cannel pile-up counters
180 PU_D = 0; PU_P = 0; PU_T = 0;
181
182 % Inter-excitation period event counter
183 j = 0;
184
185 % Create memories to store previous event information
186 prv = [ -t_P, 0 ];
187 prv_D = zeros(1,N_D) - t_D;
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
```

```
205 % Loop over all events
206 for i=1:N_events
207   % Calculate absolute time of current photon event
208   cur = per*events(2,i) + neg*events(1,i);
209   % If we're in the same excitation period
210   if ( prv(2) == events(2,i) )
211     % If we are over a SPAD dead-time from previous event
212     if ( cur - prv_D(events(3,i)) ) > t_D
213       % If we are over a SiPM pulse-width from previous event
214       if ( cur - prv(1) ) > t_P
215         % If we still have a timing channel available
216         if ( j < N_T )
217           % Add one to event counter
218           j = j + 1;
219           % Calculate which timing channel is being used
220           ch = j + N_T*mod( events(2,i), N_M );
221           % Calculate histogram bin for current event
222           bin = ceil( events(1,i) / R_T2( ch ) );
223           % Increment decay histogram
224           decay2_PU_ch(ch, bin) = decay2_PU_ch(ch, bin) + 1;
225         else
226           PU_T = PU_T + 1; % Timing pile-up
227         end
228       else
229         PU_P = PU_P + 1; % SiPM Pulse-width pile-up
230       end
231     else
232       PU_D = PU_D + 1; % Detector Dead-time pile-up
233     end
234
235
236
237
238
239
240
241
242
243
244
245
```

```
246   % Reset case, new timing window reached.
247   else
248     % Reset event counter
249     j = 0;
250     % If we are over a SPAD dead-time from previous event
251     if ( cur - prv_D(events(3,i)) ) > t_D
252       % If we are over a SiPM pulse-width from previous event
253       if ( cur - prv(1) ) > t_P
254         % Add one to event counter
255         j = j + 1;
256         % Calculate which timing channel is being used
257         ch = j + N_T*mod( events(2,i), N_M );
258         % Calculate histogram bin for current event
259         bin = ceil( events(1,i) / R_T2( ch ) );
260         % Increment decay histogram
261         decay2_PU_ch(ch, bin) = decay2_PU_ch(ch, bin) + 1;
262       else
263         PU_P = PU_P + 1; % SiPM Pulse-width pile-up
264       end
265     else
266       PU_D = PU_D + 1; % Detector Dead-time pile-up
267     end
268   end
269   % Add current event information to previous memories
270   prv = [ cur, events(2,i) ];
271   prv_D(events(3,i)) = cur;
272 end
273
274 % Create total decay from each timing channel contribution
275 if ( N_T*N_M > 1)
276   decay2_PU = sum(decay2_PU_ch);
277 else
278   decay2_PU = decay2_PU_ch;
279 end
280
281 % Convert lost events to percentages
282 PU_T = 100*PU_T / N_events;
283 PU_P = 100*PU_P / N_events;
284 PU_D = 100*PU_D / N_events;
285
286 clear events;
```

```
287 %% CMM (No Pile-Up)
288
289 % Convert peak histogram position to timing bin resolution
290 T_P2 = T_P/R_T;
291
292 % Calculate FIRST and LAST as bin position for indexing
293 % depending on whether reverse start-stop
294 if ( rev == 1 )
295   LAST = T_P2;
296   FIRST = 1;
297 else
298   LAST = 1000;
299   FIRST = T_P2 + 1;
300 end
301
302 % Calculate FIRST, LAST and m in ps for calculations
303 LAST2 = LAST*R_T;
304 FIRST2 = FIRST*R_T;
305 m = LAST - FIRST + 1;
306
307 % Need to know noise per time-bin (npb)
308 npb = ( DCR*time ) / bins;
309
310 % Calculate total number of processed events
311 cmm_count = sum( decay2(1,FIRST:LAST) );
312 % Calculate sum of histogram codes
313 cmm_sum = sum( t2(1,FIRST:LAST).*decay2(1,FIRST:LAST) );
314 % Calculate effective number of processed events (-DCR)
315 cmm_count2 = cmm_count - m*npb;
316 % Calculate effective sum of histogram codes (-DCR)
317 cmm_sum2 = cmm_sum - (( m*npb*(LAST2 + FIRST2) )/2);
318
319 % Calcualte raw CMM value
320 cmm = cmm_sum2 / cmm_count2;
321
322 % Calcualte actual CMM value, depending on reverse start-stop
323 if ( rev == 1 )
324   cmm = LAST2 - cmm;
325 else
326   cmm = cmm - FIRST2;
327 end
```

```
328 %% CMM (Pile-Up)
329
330 % Calculate total number of processed events
331 cmm_count_PU = sum( decay2_PU(1,FIRST:LAST) );
332 % Calculate sum of histogram codes
333 cmm_sum_PU = sum( t2(1,FIRST:LAST).*decay2_PU(1,FIRST:LAST) );
334 % Calculate effective number of processed events (-DCR)
335 cmm_count2_PU = cmm_count_PU - m*npb;
336 % Calculate effective sum of histogram codes (-DCR)
337 cmm_sum2_PU = cmm_sum_PU - (( m*npb*(LAST2+FIRST2) )/2);
338
339 % Calcualte raw CMM value
340 cmm_PU = cmm_sum2_PU / cmm_count2_PU;
341
342 % Calcualte actual CMM value, depending on reverse start-stop
343 if ( rev == 1 )
344   cmm_PU = LAST2 - cmm_PU;
345 else
346   cmm_PU = cmm_PU - FIRST2;
347 end
348
349
350 %% For plotting on log, remove 0s
351 decay2(decay2 == 0) = 1;
352 decay2_PU(decay2_PU == 0) = 1;
353 decay2_PU_ch(decay2_PU_ch == 0 ) = 1;
```

A.3 Single-Channel TCSPC and CMM Pile-Up Theory

Combining Equations 2.1 and 2.12 with a finite excitation period of T, gives us;

$$\tau_{\text{CMM}}(\mu;t) = \frac{\int_0^T t \cdot e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)}\ \mathrm{d}t}{\int_0^T e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)}\ \mathrm{d}t}$$

The denominator has already been calculated, as given by Equation 2.4 (excluding the $\frac{1}{T}$ term). The numerator becomes (where $\mathrm{Ei}(x)$ is the exponential integral, described by $\mathrm{Ei}(x) = \int_{-\infty}^{x} e^t/t\ \mathrm{d}t$);

$$\begin{aligned}\int_0^T t \cdot e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)}\ \mathrm{d}t &= \left[\frac{-\tau \cdot e^{-\mu}\left[\tau \cdot \mathrm{Ei}\left(\mu \cdot e^{-t/\tau}\right) + t \cdot e^{\mu \cdot e^{-t/\tau}}\right]}{\mu}\right]_0^T \\ &= \frac{\tau \cdot e^{-\mu}\left[\tau \cdot \mathrm{Ei}\left(\mu\right) - \tau \cdot \mathrm{Ei}\left(\mu \cdot e^{-T/\tau}\right) - T \cdot e^{\mu \cdot e^{-T/\tau}}\right]}{\mu}\end{aligned}$$

Therefore the fluorescence lifetime with classical pile-up by CMM can be given by;

$$\tau_{\text{CMM}}(\mu;t) = \frac{e^{-\mu}\left[\tau \cdot \mathrm{Ei}\left(\mu\right) - \tau \cdot \mathrm{Ei}\left(\mu \cdot e^{-T/\tau}\right) - T \cdot e^{\mu \cdot e^{-T/\tau}}\right]}{1 - e^{-\mu\left(1-e^{-T/\tau}\right)}}$$

For a normalised fluorescence lifetime of $\tau = 1.0$ and $T = 10 \cdot \tau$, this can be simplified to;

$$\tau_{\text{CMM}}(\mu;t) = \frac{e^{-\mu}\left[\mathrm{Ei}\left(\mu\right) - \mathrm{Ei}\left(\mu \cdot e^{-10}\right) - 10\right]}{1 - e^{-\mu}}$$

A.4 Channel Pulse-Width Pile-Up Theory

The following derivation of theoretical expressions is provided courtesy of Dr. Jochen Arlt.

Emissions of fluorescence photons can be described by a Poisson distribution. If μ is the average number of photons per laser cycle incident on the detector, then the following equation describes the probability of having n photons arriving within any given laser cycle;

$$f_\mu(n) = \frac{\mu^n \cdot e^{-\mu}}{n!}$$

While the probability of having multiple photons arriving within a signal cycle is negligible for very low μ, it generally has to be taken into account as additional photons in a given laser cycle lead to photon loss (TCSPC pile-up).

For an ideal sensor, the probability of detecting a photon which arrives as the m^{th} out of a total of n photons is given by;

$$P_{m,n}(t) = w_{m,n} \times [P_{\text{before}}(t)]^{m-1} \times P(t) \times [P_{\text{after}}(t)]^{n-m}$$

where $P_{\text{before}}(t)$, $P(t)$, $P_{\text{after}}(t)$ are the probability of a photon arriving before, at or after delay time t, respectively, and $w_{m,n}$ is the statistical weight. As the photons are indistinguishable the weight is simply given by a binomial coefficient;

$$w_{m,n} = \binom{n-1}{m-1}$$

However, for a real sensor the system is blind for a dead-time t_P, so if any photon has arrived within the time interval from max(0, $t - t_P$) to t, the m^{th} photon will no longer get detected.

The m^{th} photon will therefore only be detected with a reduced probability which can be expressed by replacing $P_{\text{before}}(t)$ with $P_{\text{before}}(t - t_P)$ (which is assumed to be 0 for $t < t_P$);

$$P_{\text{det}\,m,n}(t) = w_{m,n} \times [P_{\text{before}}(t - t_P)]^{m-1} \times P(t) \times [P_{\text{after}}(t)]^{n-m}$$

Therefore the total number of detected photon events due to the m^{th} arriving photon within a single excitation cycle is given by;

$$c_m(\mu;t) = \sum_{n=m}^{\infty} n \cdot f_n(\mu) \cdot P_{\text{det}\,m,n}(t)$$

Furthermore, the total number for all photons within a single excitation cycle is given by;

$$c_{tot}(\mu;t) = \sum_{n=m}^{n_{\text{max}}} c_m(\mu,t)$$

where n_{max} is the maximum number of photons that can be timed within a single excitation cycle.

For a single exponential decay ($I(t) = e^{-t/\tau}$) it is fairly straightforward to express most of these expressions analytically, giving;

$$\begin{aligned} P_{\text{before}}(t-t_P) = \quad & \int_0^{t-t_P} P(t')dt' \quad = \begin{cases} 0 & \text{for } t < t_P \\ 1-e^{-(t-t_P)/\tau)} & \text{for } t \geq t_P \end{cases} \\ P_{\text{after}}\,(t) = \quad & \int_t^{\infty} P(t')dt' \quad = e^{-t/\tau} \end{aligned}$$

Therefore the contribution of photons which arrive *first* at the detector is given by;

$$c_1(\mu;t) = e^{-t/\tau} \cdot \sum_{n=0}^{\infty} e^{-nt/\tau} \cdot f_\mu(n) = e^{-t/\tau} \cdot e^{-\mu\left(1-e^{-t/\tau}\right)}$$

And for a later, m^{th}, photon;

$$c_m(\mu;t) = e^{-t/\tau} \cdot \sum_{n=m}^{\infty} n \binom{n-1}{m-1} \left(1-e^{-(t-t_P)/\tau}\right)^{m-1} \cdot e^{-(n-m)t/\tau} \cdot f_\mu(n) \quad \text{for } t \geq t_P$$

Which can be re-written in a simpler form recursively, as given below;

$$c_m(\mu;t) = \frac{\mu}{m-1}(1-e^{-t/\tau} \cdot e^{t_P}) \cdot c_{m-1}(\mu;t) \qquad \text{for } t \geq t_P$$

A.5 TDC Mismatch

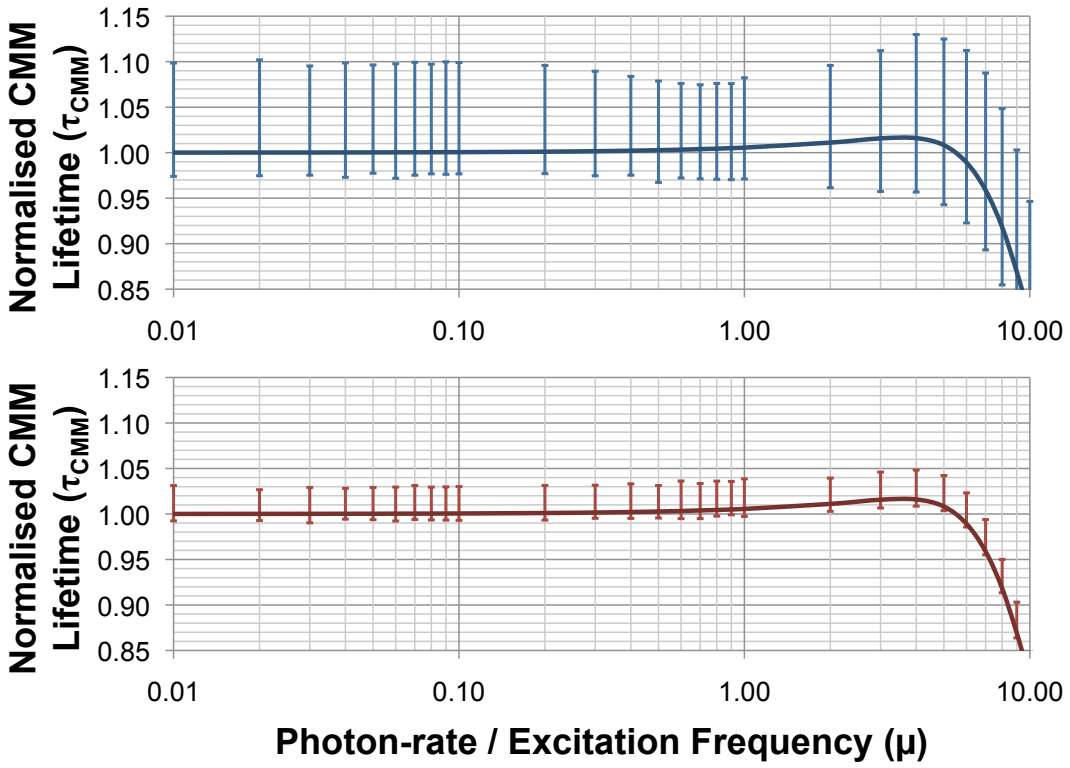


Figure A.4: *Effect of increasing μ on the worst case errors (error bars) of the CMM calculation from 100 different random TDC mismatch configurations using resetting (top) and free-running (bottom) routers. The solid lines represent the ideal CMM calculation with no TDC mismatch, from which the errors are calculated.*

A.6 *SIPM_CMM* Register Map

The order of multi-bit words should be noted carefully, as some are big-endian and some are little-endian due to routing constraints in the chip. The software register mapping takes care of this ordering.

Bits	Signal	Description
0	DIR	Control bit, to divert the serial interface to the SPAD enables, is reset separately from the rest of the serial interface memory using READ_IN. *0* – Data, Read & Clock to control serial interface. *1* – Data, Read & Clock to SPAD enables serial interface.
1	RAW_SPAD_MODE	*0* – Sampled (VALID) SPAD to ripple counter (max. 1 per laser cycle, through comparator block). *1* – Raw SPAD to ripple counter.
2	RAW_TDC_MODE	*0* – Raw TDC o/p pads disabled. *1* – Raw TDC o/p pads enabled and raw TDC values ($\times$4) to serial interface.
3	COMMS_MODE	*1* – 10-bit photon counting enabled (select bits 12:3). *0* – TDC operation enabled.
5:4	LASERSTOPSRC	*00* – Pulses in, pulses out, internal square (/2). Normal operation. *01* – Pulses in, square out (/2), internal square (/2). Chained 0. *10* – Square in, square out, internal square. Chained 1+. *11* – Selects TEST_LASER_IN (CLOCK_OUT) as laser source.
6	DELAY_ENABLE	*0* – Delay block is disabled. *1* – Delay block is enabled.
13:7	S0_DELAY	Time from STOP until MASK is asserted.
20:14	S1_DELAY	Time for GATE to stay asserted (minus S0, must be $>$ S0).
27:21	S2_DELAY	Time for MASK to stay asserted (minus S0, must be $>$ S0+S1).[1]
34:28	STOP_DELAY	Delay for STOP signal into TDC banks.
35	TDC_MODE	*0* – $\times$1 TDC bank enabled and outputs connected to serial interface. *1* – $\times$8 TDC bank enabled and outputs connected to serial interface.
36	TEST_MODE_EN	*0* – Test signals & GATE/MASK/FQUENCHN o/p pads disabled. *1* – Test signals & GATE/MASK/FQUENCHN o/p pads enabled.
37	RAW_SPAD_PAD	*0* – Raw SPAD o/p pad disabled. *1* – Raw SPAD o/p pad enabled.
38	MONOSTABLEN	*0* – SPAD monostable enabled. *1* – SPAD monostable disabled (NOTE: Cannot disable SPADs).

Table A.1: *Continued on next page*

[1] Time for FQUENCHN will be S2 – S1.

Bits	Signal	Description
39	COMP_BYPASS	*0* – Do not bypass the TDC output digital comparator.
		1 – Bypass the TDC output digital comparator.
40:42	TDC_POS_SEL	*000* – Select TDC data 9:0, giving $\approx$ 50 ps resolution.
		001 – Select TDC data 10:1, giving $\approx$ 100 ps resolution.
		...
		111 – Select TDC data 14:5, giving $\approx$ 1.6 µs resolution.
		111 – Select TDC data 15:6, giving $\approx$ 3.2 µs resolution.
43	STARTSRC	*0* – SPAD output starts TDC.
		1 – TEST_START_IN starts TDC.
44:53	CMM_LAST	Maximum TDC value for comparator (0–1024).
54:63	CMM_FIRST	Minimum TDC value for comparator (0–1024).

Table A.1: *SIPM_CMM Memory Register Map.*

A.7 Serial Interface Timing

Write

Ensuring that the *dir* bit is not currently set, 64-bits of data (*DATA*) is clocked (*CLK*) into the shift register serial interface. The data is then written to the control registers by strobing *WRITE* once.

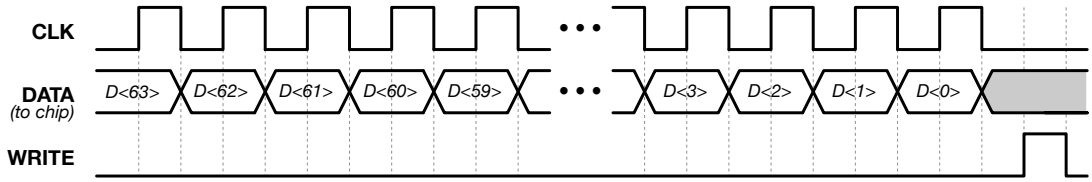


Figure A.5: *Timing diagram for a register write.*

SPAD Enables

The *dir* bit is set within the control registers by performing a serial interface write, as described above. Then, 1024-bits of data (*DATA*) is clocked (*CLK*) into the extended shift register serial interface. The data is not written in the case of SPAD enables, however the *dir* bit is reset by strobing *READ* once.

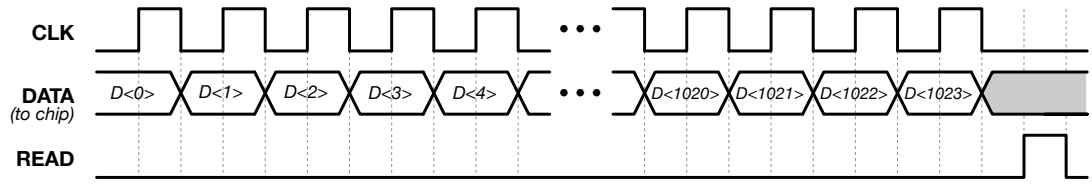


Figure A.6: *Timing diagram for setting SPAD enables.*

Read

The *READ* input is held high for a long enough duration to allow the CMM pre-calculation to settle. To ensure the data can settle, *READ* is used internally as a *GATE* to stop new TCSPC timestamps being processed. Strobing *CLK* high when the data has settled and whilst *READ* is still high will load the CMM data into the serial interface shift register. A short internal $\overline{RESET}$ signal is created using the NAND of *CLK* and *READ* to reset the CMM processing block. The data (*DATA*) can then be clocked (*CLK*) out of the device.

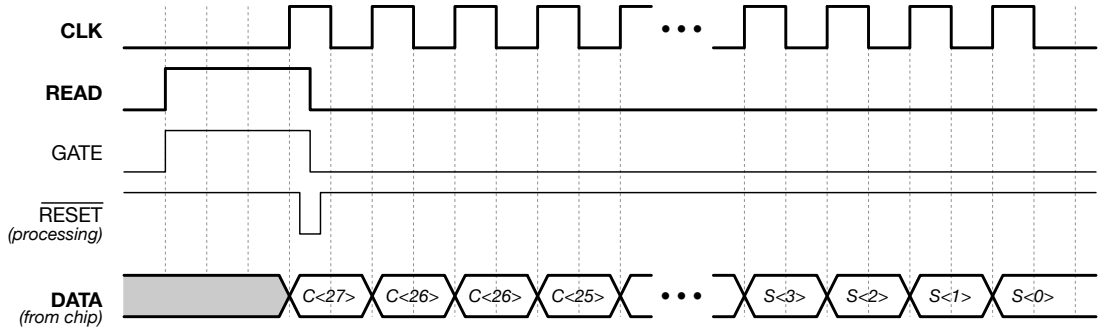


Figure A.7: *Timing diagram for a read.*

System Reset

Due to the limited number of I/O pads available, a system reset ($\overline{RESET}$) is created from the NAND of *READ* and *WRITE*. To ensure that data is not written during this reset, *WRITE* is gated when *READ* is high. To ensure no glitches, *READ* must completely overlap *WRITE*.

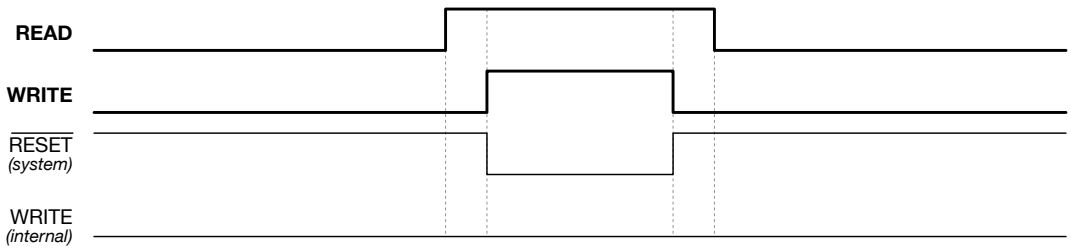


Figure A.8: *Timing diagram for a system reset.*

A.8 SIPM_CMM Padlist

The device has 38 functional pads, positioned as shown in Figure A.9 below, where the grey signals are optional test signals. Information on each pad is detailed in the Table A.2, where the pads are numbered as per their position in the 48CLCC package.

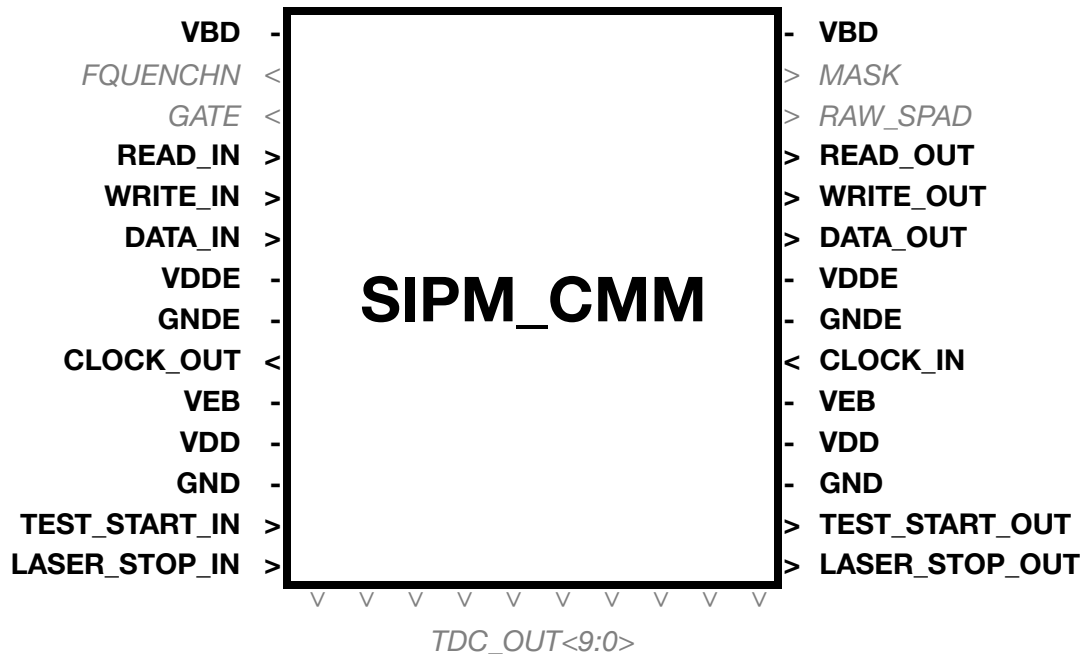
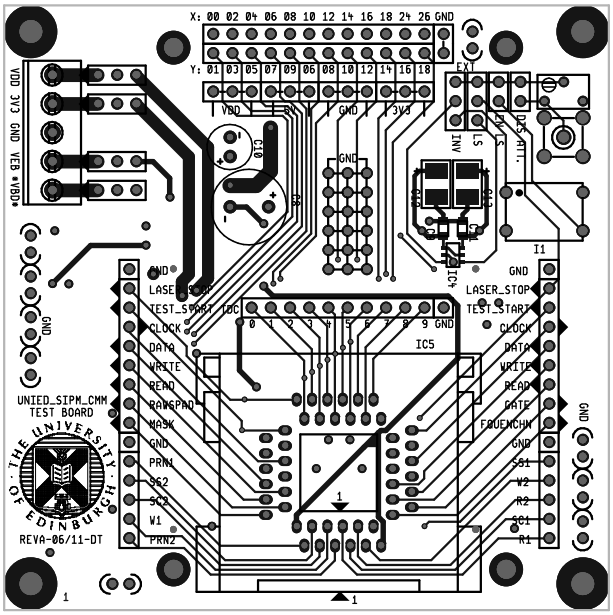


Figure A.9: *Padring signal locations in a ×2 device network.*

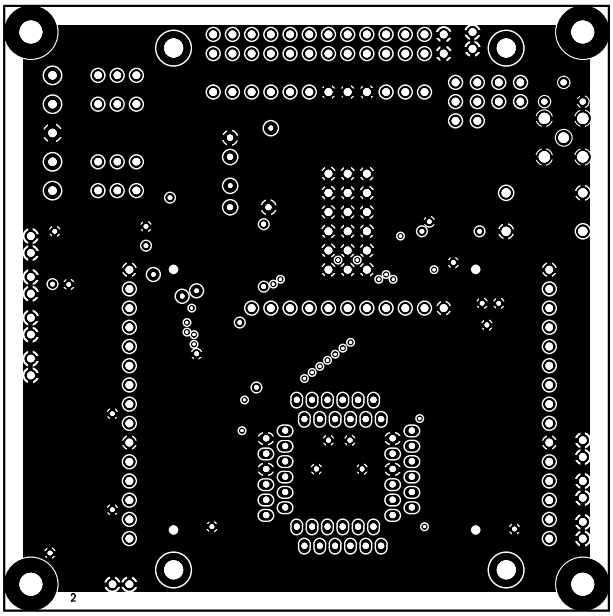
Pad	Signal	Bond	Type	IMG175 Library
6	VBD *(in)*	F/S/C	Power	ANA_TC_NOPROT (unprotected pad)
7	FQUENCHN	F	Output	BD4SCRPROG54M108_TC_FS
8	GATE	F	Output	BD4SCRPROG54M108_TC_FS
9	READ_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
10	WRITE_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
11	DATA_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
12	VDDE *(in)*	F/S/C	Power	VDDIO_3V3_65
13	GNDE *(in)*	F/S/C	Power	VSSIO_3V3_65
14	CLOCK_OUT	F/C	Output	BD4SCRPROG54M108_TC_FS
15	VEB *(in)*	F/S/C	Power	VDDCO_3V3_65
16	VDD *(in)*	F/S/C	Power	VDDIOCO_65
17	GND *(in)*	F/S/C	Power	VSSIOCO_65
18	TEST_START_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
19	LASER_STOP_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
20:29	TDC_OUT<9:0>	F	Outputs	BD4SCRPROG54M108_TC_FS
30	LASER_STOP_OUT	F/C	Output	BD4SCRPROG54M108_TC_FS
31	TEST_START_OUT	F/C	Output	BD4SCRPROG54M108_TC_FS
32	GND *(out)*	F/C	Power	VSSIOCO_65
33	VDD *(out)*	F/C	Power	VDDIOCO_65
34	VEB *(out)*	F/C	Power	VDDCO_3V3_65
35	CLOCK_IN	F/S/C	Input	BD4SCRPROG54M108_TC_FS
36	GNDE *(out)*	F/C	Power	VSSIO_3V3_65
37	VDDE *(out)*	F/C	Power	VDDIO_3V3_65
38	DATA_OUT	F/S/C	Output	BD4SCRPROG54M108_TC_FS
39	WRITE_OUT	F/C	Output	BD4SCRPROG54M108_TC_FS
40	READ_OUT	F/C	Output	BD4SCRPROG54M108_TC_FS
41	RAW_SPAD	F/S	Output	BD4SCRPROG54M108_TC_FS
42	MASK	F	Output	BD4SCRPROG54M108_TC_FS
43	VBD *(out)*	F/C	Power	ANA_TC_NOPROT (unprotected pad)

Table A.2: *Pad list.*

A.9 Evaluation Platform PCB

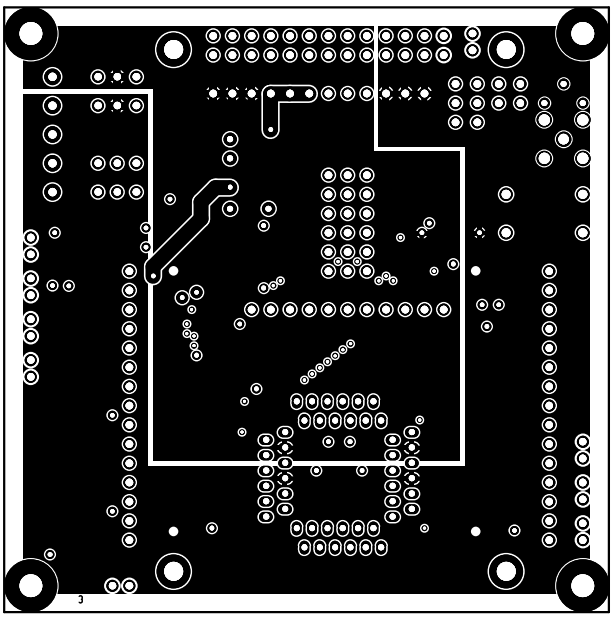


(a)

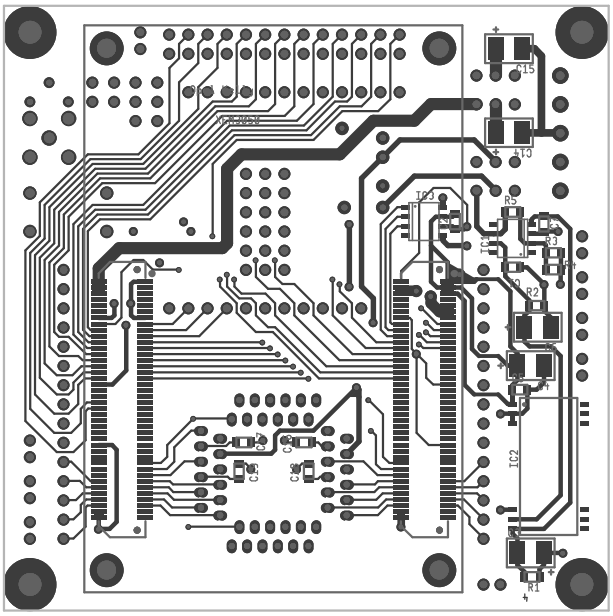


(b)

Figure A.10: (a) *Top (layer 1) copper & silkscreen and* **(b)** *inside layer 2 copper (GND).*

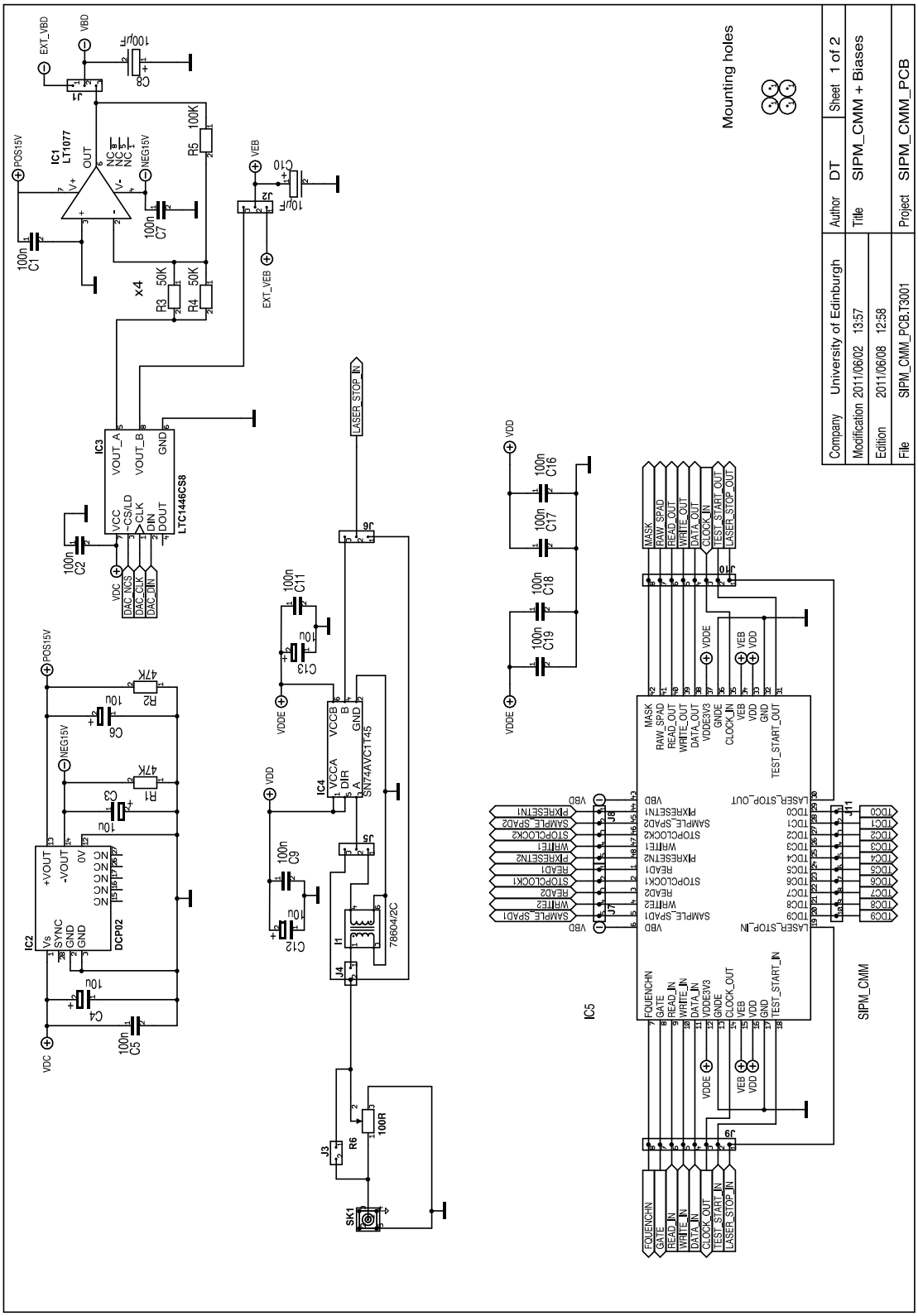


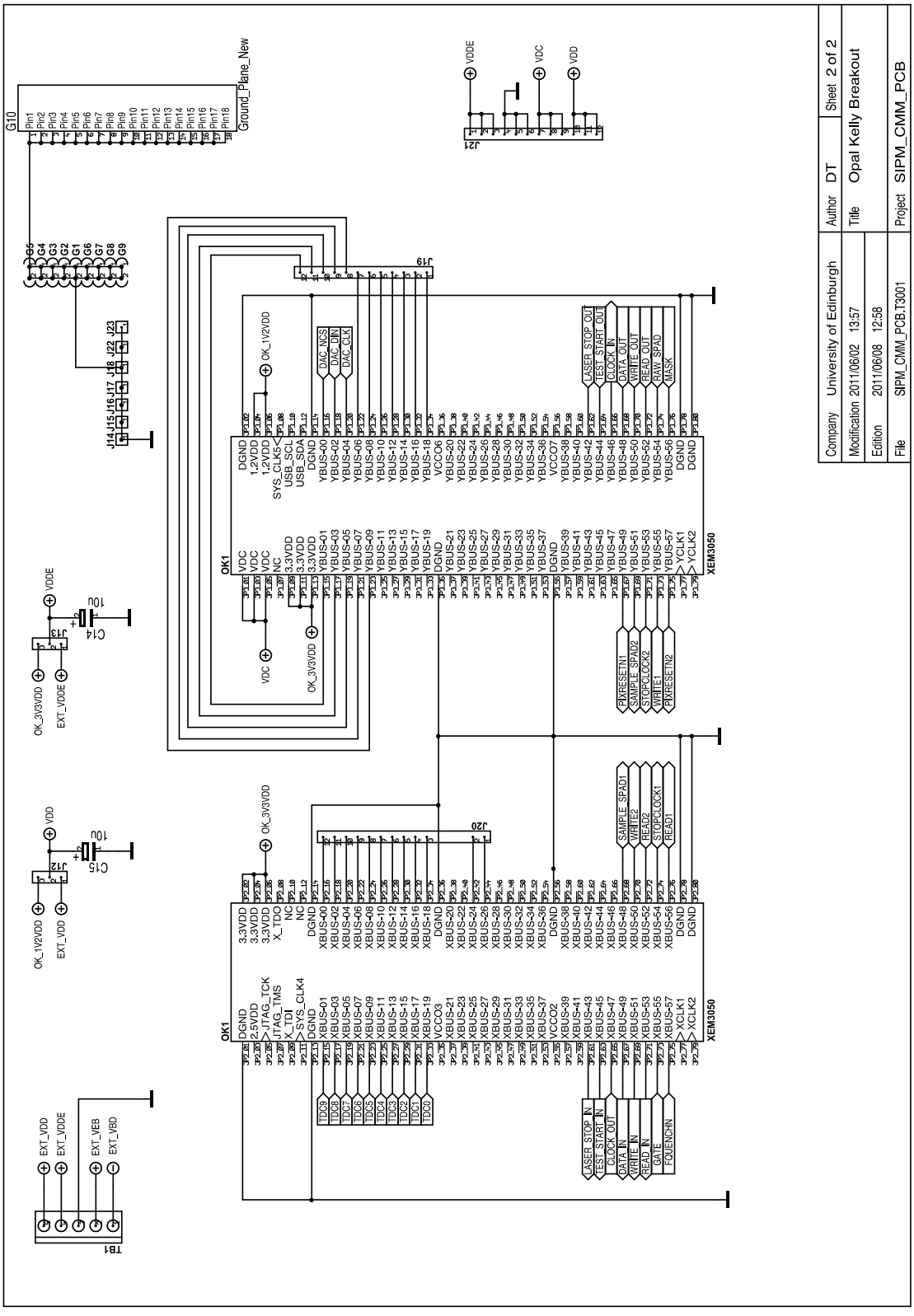
(a)



(b)

Figure A.11: (a) *Inside layer 3 copper (V_{DD} / V_{DDE}) and* **(b)** *bottom (layer 4) copper & silkscreen.*


Company University of Edinburgh
Author DT
Sheet 1 of 2
Modification 2011/06/02 13:57
Title SIPM_CMM + Biases
Edition 2011/06/08 12:58
File SIPM_CMM_PCB.T3001
Project SIPM_CMM_PCB
Mounting holes
IC1 LT1077
IC2 DCP02
IC3 LTC1446CS8
IC4 SN74AVC1T45
IC5
SIPM_CMM
EXT_VBD
VBD
EXT_VEB
VEB
POS15V
NEG15V
VDC
VDDE
VDD
LASER_STOP_IN
DAC_NCS
DAC_CLK
DAC_DIN
FOUENCHN
GATE
READ_IN
WRITE_IN
DATA_IN
CLOCK_OUT
TEST_START_IN
LASER_STOP_IN
MASK
RAW_SPAD
READ_OUT
WRITE_OUT
DATA_OUT
CLOCK_IN
TEST_START_OUT
LASER_STOP_OUT
SAMPLE_SPAD1
WRITE1
READ1
PIXRESET1
STOPCLOCK1
SAMPLE_SPAD2
WRITE2
READ2
PIXRESET2
STOPCLOCK2
TDC0
TDC1
TDC2
TDC3
TDC4
TDC5
TDC6
TDC7
TDC8
TDC9
7860#2C


Company University of Edinburgh
Author DT
Sheet 2 of 2
Title Opal Kelly Breakout
Modification 2011/06/02 13:57
Edition 2011/06/08 12:58
File SIPM_CMM_PCB.T3001
Project SIPM_CMM_PCB
XEM3050
OK1
G10
Ground_Plane_New
J19
J20
J21
TB1
C14
C15
10u
VDDE
VDC
VDD
EXT_VDD
EXT_VDDE
EXT_VEB
EXT_VBD
OK_3V3VDD
OK_1V2VDD
DAC_NCS
DAC_DIN
DAC_CLK
LASER_STOP_OUT
TEST_START_OUT
CLOCK_IN
DATA_OUT
WRITE_OUT
READ_OUT
RAW_SPAD
MASK
PIXRESETN1
SAMPLE_SPAD2
STOPCLOCK2
WRITE1
PIXRESETN2
SAMPLE_SPAD1
WRITE2
READ2
STOPCLOCK1
READ1
LASER_STOP_IN
TEST_START_IN
CLOCK_OUT
DATA_IN
WRITE_IN
READ_IN
GATE
FQUENCHN

B
PUBLICATIONS

The following journal and conference publications arose from the work undertaken throughout the entirety of my Ph.D. study and are presented in reverse chronological order. Those directly relevant to the research documented in this thesis are highlighted in **bold** and provided in full in the remainder of the Appendix.

Primary Author

[1] J. Arlt, D. Tyndall[1], B. R. Rae, D. D.-U. Li, J. A. Richardson, and R. K. Henderson, "A Study of Pile-up in Integrated Time-Correlated Single Photon Counting Systems," *Review of Scientific Instruments*, vol. 84, no. 10, Oct. 2013.

[2] D. Tyndall, B. R. Rae, D. D.-U. Li, J. Arlt, A. Johnston, J. A. Richardson, and R. K. Henderson, "A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13 µm CMOS," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 6, pp. 562–570, Dec. 2012.

[3] D. Tyndall, B. R. Rae, D.-U. Li, J. A. Richardson, J. Arlt and R. K. Henderson, "A 100Mphoton/s Time-Resolved Mini-Silicon Photomultiplier with On-Chip Fluorescence Lifetime Estimation in 0.13 µm CMOS Imaging Technology," in *International Solid-State Circuits Conference (ISSCC)*, Feb. 2012, pp. 122–123.

[4] D. Tyndall, R. J. Walker, K. Nguyen, R. Galland, J. Gao, I. Wang, M. Kloster, A. Delon and R. K. Henderson; "Automatic laser alignment for multifocal microscopy using a LCOS SLM and a 32×32 pixel CMOS SPAD array," in *Proceedings of SPIE*, vol. 8086, 2011.

[1] J. Arlt and D. Tyndall contributed equally to this work.

Co-Author

[5] M. Kloster-Landsberg, D. Tyndall, I. Wang, R. J. Walker, R. K. Henderson and A. Delon; "Note: Multi-confocal fluorescence correlation spectroscopy in living cells using a complementary metal oxide semiconductor-single photon avalanche diode array," *Review of Scientific Instruments*, vol. 84, no. 7, Jul. 2013.

[6] D. D.-U. Li, S. Poland, S. Coelho, D. Tyndall, W. Zhang, R. Walker, J. Richardson, and R. Henderson, "Advanced Fluorescence Lifetime Imaging Algorithms for CMOS Single-Photon Sensor Based Multi-focal Multi-photon Microscopy," in *Engineering in Medicine and Biology Conference (EMBC)*, Osaka, Jul. 2013.

[7] S. Poland, S. Coelho, N. Krstajic, D. Tyndall, R. J. Walker, J. Monypenny, D. D.-U. Li, R. K. Henderson, and S. Ameer-Beg, "Development of a Fast TCSPC FLIM-FRET Imaging System," in *Proceedings of SPIE*, vol. 8588, San Francisco, Feb. 2013.

[8] S. Coelho, S. Poland, N. Krstajic, D. D.-U. Li, J. Monypenny, R. J. Walker, D. Tyndall, T. C. Ng, R. K. Henderson, and S. Ameer-Beg, "Multibeam multiphoton microscopy with adaptive optical correction," in *Proceedings of SPIE*, vol. 8588, San Francisco, Feb. 2013.

[9] D. D.-U. Li, S. Ameer-Beg, J. Arlt, D. Tyndall, R. J. Walker, D. R. Matthews, V. Visitkul, J. A. Richardson, and R. K. Henderson, "Time-Domain Fluorescence Lifetime Imaging Techniques Suitable for Solid-State Imaging Sensor Arrays," *Sensors*, vol. 12, no. 5, pp. 5650–5669, May 2012.

[10] D. D.-U. Li, J. Arlt, D. Tyndall, R. J. Walker, J. A. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm," *Journal of Biomedical Optics*, vol. 16, no. 9, p. 096012, Sep. 2011.

[11] G. Giraud, H. Schulze, D. D.-U. Li, T. T. Bachmann, J. Crain, D. Tyndall, J. A. Richardson, R. J. Walker, D. Stoppa, E. Charbon, R. K. Henderson, and J. Arlt, "Fluorescence lifetime biosensing with DNA microarrays and a CMOS-SPAD imager," *Biomedical Optics Express*, vol. 1, no. 5, pp. 1302–1308, Jan. 2010.

B.1 Arlt, Tyndall et. al., Rev. Sci. Inst., 2013

REVIEW OF SCIENTIFIC INSTRUMENTS **84**, 103105 (2013)

A study of pile-up in integrated time-correlated single photon counting systems

Jochen Arlt,[1] David Tyndall,[2,3] Bruce R. Rae,[4] David D.-U. Li,[5] Justin A. Richardson,[3] and Robert K. Henderson[2]
[1]*SUPA, COSMIC, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom*
[2]*Institute for Integrated Micro and Nano Systems, School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, United Kingdom*
[3]*Dialog Semiconductor, 2 Multrees Walk, Edinburgh EH1 3DQ, United Kingdom*
[4]*STMicroelectronics, Pinkhill, Edinburgh EH12 7BF, United Kingdom*
[5]*Department of Engineering and Design, School of Engineering and Informatics, University of Sussex, Brighton BN1 9QT, United Kingdom*

(Received 7 August 2013; accepted 20 September 2013; published online 10 October 2013)

Recent demonstration of highly integrated, solid-state, time-correlated single photon counting (TCSPC) systems in CMOS technology is set to provide significant increases in performance over existing bulky, expensive hardware. Arrays of single photon single photon avalanche diode (SPAD) detectors, timing channels, and signal processing can be integrated on a single silicon chip with a degree of parallelism and computational speed that is unattainable by discrete photomultiplier tube and photon counting card solutions. New multi-channel, multi-detector TCSPC sensor architectures with greatly enhanced throughput due to minimal detector transit (dead) time or timing channel dead time are now feasible. In this paper, we study the potential for future integrated, solid-state TCSPC sensors to exceed the photon pile-up limit through analytic formula and simulation. The results are validated using a 10% fill factor SPAD array and an 8-channel, 52 ps resolution time-to-digital conversion architecture with embedded lifetime estimation. It is demonstrated that pile-up insensitive acquisition is attainable at greater than 10 times the pulse repetition rate providing over 60 dB of extended dynamic range to the TCSPC technique. Our results predict future CMOS TCSPC sensors capable of live-cell transient observations in confocal scanning microscopy, improved resolution of near-infrared optical tomography systems, and fluorescence lifetime activated cell sorting. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4824196]

I. INTRODUCTION

Time-correlated single photon counting (TCSPC) is the favoured technique to characterise fluorescence decay lifetimes of low light signals emitted from single dye molecules in response to synchronized optical impulses. It provides picosecond accuracy and high photon efficiency, enabling precise multi-exponential studies of molecular interactions within live cells using techniques such as Forster Resonant Energy Transfer (FRET).[1] TCSPC is conventionally implemented using a pulsed excitation source, a discrete detector such as an avalanche photodiode (APD) or photomultiplier tube (PMT), external time-to-digital conversion (TDC) hardware, and a PC to compute the decay constant. Pixel-by-pixel data acquisition of TCSPC in conjunction with a confocal scanning microscope setup allows 3D sectioning of samples and Fluorescence Lifetime Imaging (FLIM). Generally, the low fluorophore concentrations in cellular FRET measurements or autofluorescence imaging and sample photostability limit the available count rate to well within the few megahertz counting capability of typical hardware.[2] In the best case, the low photon emission allows acquisition of a 10 kilopixel FLIM image in approximately 1 s. However, certain FLIM experiments can be run at a much higher count rate such as chlorophyll transients or yeast autofluorescence.[2] Furthermore, other non-imaging applications of TCSPC such as flow cytometry, fluorescence activated cell sorting, high throughput screening (HTS), and functional near infrared spectroscopy (fNIRS) require acquisition at peak photon rates in the 10–100 MHz range.[1]

TCSPC count rates are limited by three effects: classical TCSPC pile-up, where only one photon event can be measured per excitation period;[1,3,4] counting loss due to the dead-time of the timing electronics[5] (typically greater than 100 ns); and counting loss due to pulse overlap caused by detector dead-time. The classic TCSPC pile-up effect requires photon count rates to operate at a maximum of 10%–20% of the excitation repetition rate[1] accepting a few percent error in lifetime estimates. However, with modern megahertz excitation rates, the counting loss due to the dead-time of the timing electronics reduces this to only 1%–5%. These losses have been addressed by increasing the excitation repetition rate to 100 MHz at the expense of only being able to measure sub-nanosecond lifetimes[6] and by operating up to 8 parallel timing channels to provide a total photon count rate of 50 MHz.[2] Multi-detector arrangements can alleviate pile-up due to detector dead time; for a single PMT, the practical limit is of the order of 5–10 MHz.[1] Often the photon events from an array of detectors are multiplexed by a router through a single timing channel, limited by routing channel and timing system dead time to 10 MHz count rate. In multi-detector

systems, pile-up can be suppressed by inhibiting the timing system on detection of several photons occurring within the same signal period in different detector channels,[7,8] however, such an approach is photon inefficient. In general, TCSPC operation above 10 MHz count rate requires multi-module TCSPC systems comprising both parallel detector and timing channels. However, the physical system size, cost, and processing requirements limit this approach to a small number of channels.[1]

An integrated solid-state CMOS implementation of a multi-module TCSPC system has recently been demonstrated.[9] The high speed and compactness of electronics implemented in modern very large scale integration (VLSI) radically alter the constraints on TCSPC system design and related single photon techniques.[10] For example, large numbers of single photon avalanche diode (SPAD) detectors and compact time to digital converter (TDC) timing channels can be integrated on a single chip. Indeed, a 20k multi-channel TCSPC imager has been presented recently.[11] Moreover, high speed digital circuit techniques such as pulse shortening and pipelined converter operation allow detector and timing channel dead time to be reduced by several orders of magnitude over discrete hardware, from a few 100 ns to few 100 ps duration. CMOS SPAD detectors offer extremely high photon detection efficiencies (>50%), low dark count (<100 Hz), and after-pulsing rates (<0.1%) at high fill-factors (>70%). Nevertheless, they suffer from relatively long dead time (≈10–50 ns) compared to the single photon response of PMTs (≈1 ns), with potential for significant detector pile-up effects. Finally, on-chip TCSPC histogramming or even direct signal processing to compute lifetime is feasible to limit data transfer and memory requirements of field-programmable gate array (FPGA) based systems. Thus, the latency with which the lifetime estimates can be obtained can be greatly reduced (few 10s of microseconds).

The study presented in this paper aims to explore the new freedom to design TCSPC systems with large numbers of parallel channels, low latency, and dead time. We seek to determine the ultimate performance of multi-module VLSI TCSPC implementations through mathematical modeling and computer simulation. In Sec. II, we present possible architectures for a multi-channel TCSPC chip and develop models for photon loss and pile-up. In Sec. III, we examine the effect of pulse-shortening, SPAD dead time, and number of channels on photon-pile-up distortion and ultimate photon throughput. In Sec. IV, we compare the predictions of our model applied to our chip implementation[9] with measured experimental results.

II. PILE-UP IN INTEGRATED TCSPC SENSORS

In TCSPC systems, not every single photon arriving at the detector contributes to the measured signal. Quantum efficiency and fill factor determine the detection efficiency of the device, and although they are crucial for the overall photon sensitivity, the associated photon loss has no direct effect on the lifetime measurement as it affects all photons with the same probability. However, there are also signal losses which depend on the photon arrival time and can lead to inaccurate estimation of the lifetime ("Pile-up"). Classical pile-up, which is due to the fact that traditional TCSPC systems can only time-stamp at most one photon per signal cycle, is probably the best-known of such distortions and has been discussed in great detail in literature. For example, Becker's book[1] covers classical pile-up together with most of the other effects which distort the recorded arrival time distribution. These can be roughly divided into 3 categories as illustrated in Fig. 1:

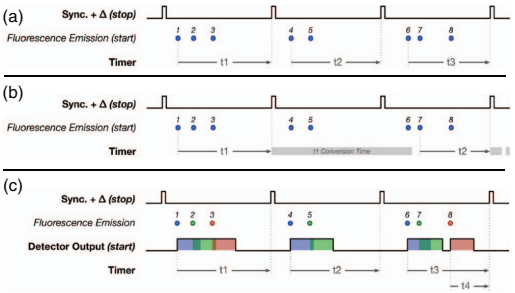


FIG. 1. (a) Classical TCSPC, (b) timing dead-time, and (c) detector dead-time TCSPC pile-up.

(a) Classical TCSPC pile-up: The timing hardware is unable to process more than one photon event in any given excitation period.
(b) Timing dead-time pile-up: The timing hardware requires a certain time t_T (typically > 100 ns) after receiving a photon to process the event and produce a timestamp. During this time, it is unable to process any further photon events.
(c) Detector dead-time pile-up: A detector registering a photon is unable to detect further photons for a dead time t_D.

A classical TCSPC system consists of a single detector combined with a single timer element, but for a sensor with more elements there is a choice between several architectures (Fig. 2). Generally, the optical signal is spread over N_D independent detectors. For systems based on photomultiplier systems, the size and cost of hardware limits N_D, however for integrated CMOS, many thousand SPADs can be employed in a silicon photomultiplier (SiPM) like arrangement. The simplest TCSPC system couples a single detector into a single timing channel, and multi-module systems comprise multiple such arrangements (Arch. I). More generally, the pulses representing detected photons can be distributed via N_C channels to N_T timing measurement blocks. As before, where discrete TCSPC cards are employed N_T is restricted to low integers due to size and cost, however in integrated form, N_T may easily exceed several thousand. Where N_D is not equal to N_T, some multiplexing system is necessary to combine detector outputs and redistribute them to timing measurement blocks. In a conventional TCSPC system, this is accomplished by a router consisting of a summing amplifier combining detector pulses and an encoder conveying digital channel number information.[1] In CMOS technology, the router can be implemented entirely by digital electronics operating on the logic level pulses from the SPADs. The summing amplifier

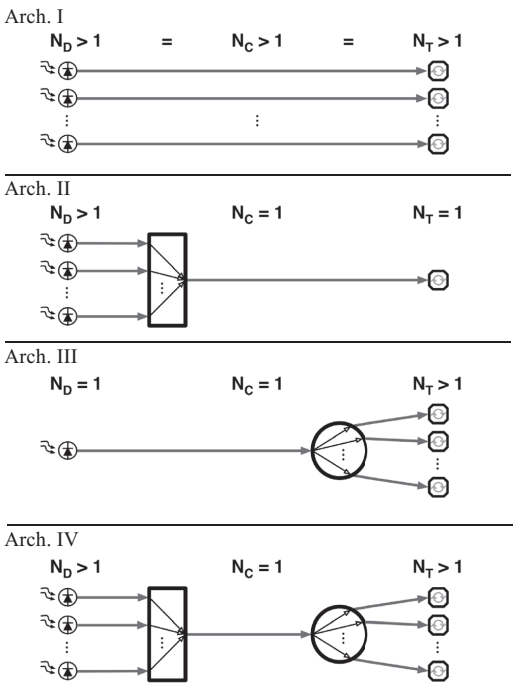


FIG. 2. Possible (sub-)sensor architectures to combine N_D parallel detectors and N_T timers.

becomes an OR gate and the router may employ sophisticated schemes to distribute the summed digital pulses to the N_T available timing measurement blocks. Therefore, a final pile-up mechanism can be distinguished due to this summation:

(d) Router pile-up : The loss of detected photons within the router due to the summation of pulses whereby detector pulses occurring within a time t_P of one another will coalesce into a single pulse.

As SPAD pulses are ≈50 ns in duration, router pile-up starts to become important at around a few MHz pulse rates. However, in CMOS implementation the router may also perform pulse shaping on the detector pulses, shortening their duration to t_P. This is readily accomplished by simple pulse-shortening circuits allowing t_P to approach a few logic gate delays in modern CMOS (≈100 ps). This step is clearly of less importance for the few nanosecond output of PMTs which is perhaps why this effect, and that of detector pile-up, are little discussed in the literature. An integrated CMOS sensor has been presented recently[9] and its specific architecture is shown in Figure 3. This device's parameters provide the basis for the modeling and analysis that follow in Sec. III.

TABLE I. Parameters of device explored in the numerical model as well as for the experimental data.

Variable	Description	Modeled value(s)	Values in experiment
N_D	Number of SPAD detectors in SiPM	1–1024	8
t_D	SPAD dead-time	10–50 ns	10 ns
DCR	Dark Count Rate (DCR) per SPAD	0–1 kHz	700 Hz
t_P	SiPM output pulse-width	< 1 ns	550 ps
N_T	Number of parallel timing channels	1–100	8
R_T	Resolution of timing-channels	50–400 ps	113 ps

III. PILE-UP MODEL AND THEORY

In this section, we will study the detector, router, and timer pile-up effects of a generic, integrated TCSPC system (Fig. 2) through MATLAB modeling and present analytical expressions for the histogram distortion introduced by "router pile-up." Typical parameter ranges of integrated TCSPC realizations from Table I will be used in this study. The accuracy of our model is confirmed with measured results from a CMOS silicon photomultiplier[9] in Sec. IV.

A. Numerical model

To study the behaviour of different integrated sensor designs, we have numerically modeled their performance in MATLAB, taking into account all of the sources of histogram distortions discussed in Sec. II. The MATLAB random number generator is used to simulate individual photon events with the appropriate statistical distribution, i.e., a single exponential distribution as a function of delay time with lifetime τ on an uncorrelated background due to the detector dark count rate (DCR) and a Poisson distribution into individual

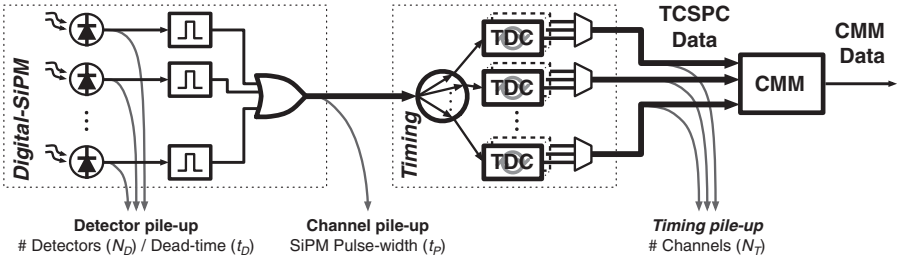


FIG. 3. Specific sensor architecture of our hardware implementation, corresponding to Arch. IV.

excitation cycles and parallel detectors as a function of μ (number of photons per excitation cycle). These random photon events are processed by code that simulates the various photon loss mechanism of the sensor, as shown schematically in Figure 1, to determine which events will get detected and timed. Detection rate and arrival time histograms can be generated to study the effects of the separate loss mechanisms individually as well as their combined effects.

Furthermore, the MATLAB code calculates the centre-of-mass[12] (or first moment) of the decay histogram to estimate its single-exponential lifetime τ, as this is also directly performed in our hardware realization (see Sec. IV A)

$$\tau_{CMM} \approx \frac{\int_0^T t\exp(-t/\tau)dt}{\int_0^T \exp(-t/\tau)dt} \approx \left(\frac{\sum_{j=0}^{M-1} jN_j}{N_c} + \frac{1}{2}\right) R_T, \tag{1}$$

where T is the duration of the temporal measurement window which is quantized into M time bins of width R_T. N_j is the number of recorded counts in the jth time bin ($j = 0, 1, \ldots, M-1$), and N_c is the total signal count within the measurement window. Full details of the MATLAB implementation can be found elsewhere.[13]

Each of the three pile-up mechanisms will first be studied individually, beginning with effects of the detectors parameters (N_D, t_D, and DCR), then the number of timing channels (N_T) and finally the router pile-up. When looking at each of these aspects, the parameters not being studied are idealised (N_T, $N_D = \infty$ and t_P, t_D, DCR $= 0$). Finally, the effects of all parameters combined will be investigated.

Each of these investigations will follow a similar strategy. To begin with, representative histograms are captured for fixed values of μ, to highlight how each form of pile-up distorts the decay in different ways. Then photon loss data and lifetime estimates are presented as a function of the number of photons per excitation cycle μ. Unsurprisingly, it is found that distortions become more apparent for high μ but can be counter-acted by increasing the number of detection elements and/or timers. To highlight the specific requirements for different sensor designs, we determine the number of elements needed to at a given rate μ to keep the centre-of-mass method (CMM) lifetime estimate within 1% of the true value.

Typically, the absolute lifetime value of the decay is irrelevant, but its relative value in terms of R_T, f_E, t_D, and t_P can be important. For the purposes of consistency, the results shown in the following use $R_T = \tau/100$ and $\tau = 0.1/f_E$, while t_D and t_P are varied.

B. Detector pile-up

In order to be able to detect high photon rates, the use of separate detection elements is needed. By distributing photons evenly over a large number of detection elements, the probability of photons arriving at the same element during the same signal cycle can be minimised. Performing simulations for an increasing number N_D of detectors with dead time t_D and dark count rate DCR while assuming no distortions from the router and timers (making it equivalent to Arch. II), it is found that the shape of the distortion depends strongly on the duration of dead time t_D relative to the lifetime τ (Fig. 4). If

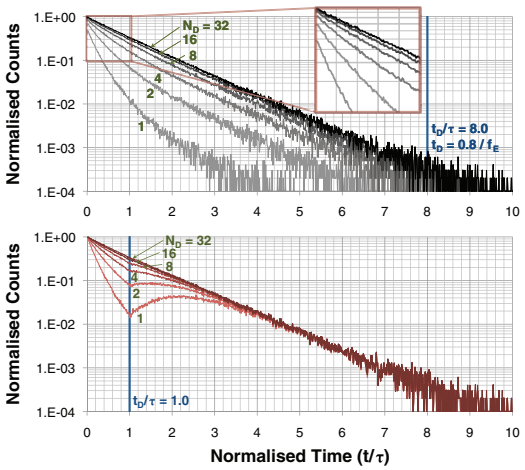


FIG. 4. Effect of increasing N_D on the captured histogram for $t_D/\tau = 8.0$ (top) and $t_D/\tau = 1.0$ (bottom), $\mu = 5.0$ and DCR of 1 KHz.

the dead time is comparable to the lifetime, the initial part of the decay is distorted similar to classical pile-up but for delay times exceeding t_D most photons can be detected and timed. However, long dead times typical for SPAD detectors often exceed the lifetimes of common fluorophores and can even be comparable to the duration of an excitation period $1/f_E$. This leads to cyclic effects, as detectors might not recover fast enough to detect photons arriving early in the signal period if they detected an event in the preceding cycle. For long dead times, each detector can effectively only detect a single event per cycle, whereas shorter dead times mean that individual elements can detect multiple events.

In either case, the distortions for a given photon rate μ become less severe as the number of detectors N_D is increased and fewer events are lost (Fig. 5). This also improves the

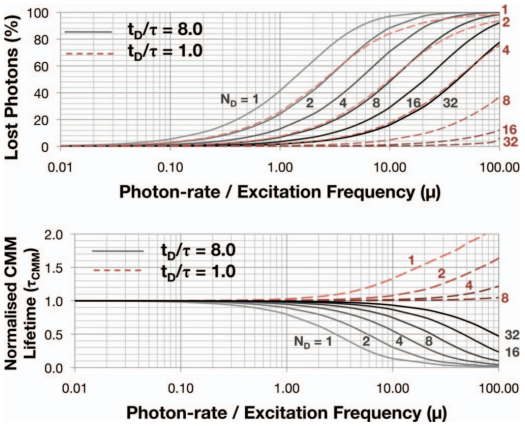


FIG. 5. (Top) Photons lost due to detector pile-up and (bottom) lifetime estimates for varying number of detection elements (N_D) and dead-times (t_D/τ) of 8.0 (solid) and 1.0 (dashed).

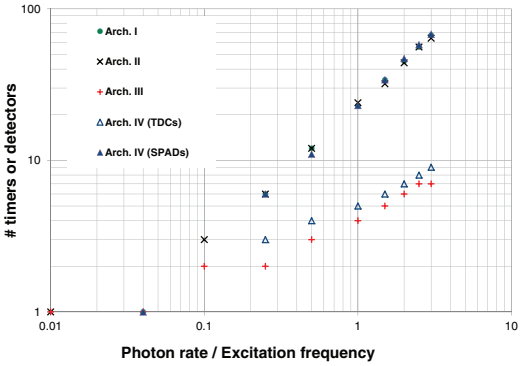


FIG. 6. Minimum number of timers and detectors needed to achieve a CMM lifetime estimate with errors below 1% for different sensor architectures introduced in Fig. 2: Arch. I: Parallel channels of single detector and timer (•). Arch. II: Parallel detectors with ideal timers (×). Arch. III: Ideal detectors with parallel timers (+). Arch. IV: "Routed" sensor, but without router pile-up (△: timers, ▲: detectors).

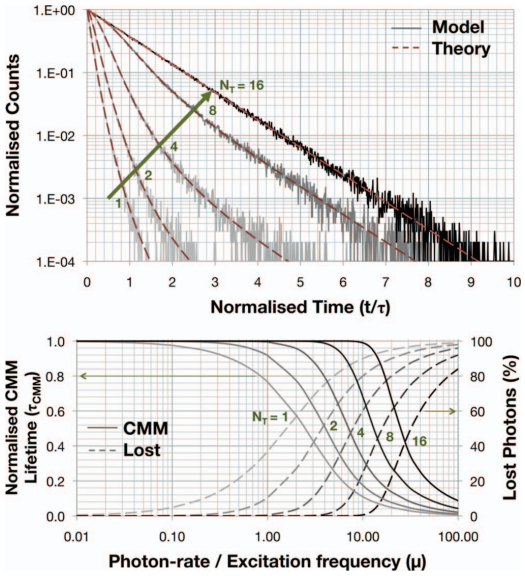


FIG. 7. (Top) Modeled arrival time histograms compared to the analytical prediction for a high photon rate of $\mu = 10.0$ and (bottom) percentage of lost photons and normalised lifetime estimates for an increasing number of timing channels N_T.

lifetime estimates based on the centre of mass algorithm which are also shown as part of this figure. For short detector dead times, the lifetime is typically overestimated but converges fairly quickly to the proper values as the number of detectors is increased. However, for detectors with long dead time the lifetime is underestimated (as in classical pile-up) and many more detectors are needed to achieve a good estimation at high photon rates.

These data can be used to find the minimum number of detectors required to obtain a predefined acceptable lifetime error at a given photon rate, e.g., within 1% of its true value (Fig. 6). It is found that for long dead times the number of detectors needed to stay below an acceptable lifetime error increases proportional to the incident photon rate. That is, the number of detectors has to be large enough to keep the count rate for *each* of the detectors below a maximum value $\mu_{D,1\%} = 4.5\%$. Note that the dark count rate has a negligible effect at the high signal rates considered here, as even a fairly high DCR of 1 kHz per SPAD is irrelevant at total count rates of $\mu f_E \gtrsim 10$ MHz.

C. Timer pile-up

Next, we investigate the effect of having a number of parallel timing channels N_T available per excitation period, assuming no pile-up distortions from any other sensor component (Arch. III). In this case, the mth photon arrival event within each excitation period is routed to the mth timing channel and all of the first N_T detected photons will also be timed.

Figure 7(a) shows the effect of varying the number of timing channels available per excitation period on the resulting TCSPC histograms for a very high photon-rate ($\mu = 10$), using both the model (solid) and analytical predictions (dashed) created by summing c_m for $m = 1 \rightarrow N_T$ for each N_T (see the Appendix, Eqs. (A5) and (A7) with $t_p = 0$). If only a small number of timers is available, many events are not timed and the histograms are distorted in a very similar way to "classical" pile-up. But as the number of timers increases, the number of lost photons and the resulting histogram distortion decreases very quickly. Again, this is reflected in the lifetime estimation.

When looking at the number of timers needed to obtain a predefined acceptable lifetime error for a given photon rate, it is found that it increases slower than proportionally (Fig. 6). Strikingly, in the sensor architecture with "routed" timers (Arch. III), a lot fewer timing elements are needed to achieve a lifetime estimate of the same quality than SPAD detector elements when using architecture II. This different scaling might at first sight appear surprising, as both the individual detectors and timers considered here can only process at most one photon per cycle. However, photon events are distributed randomly amongst the N_D detectors, whereas the distribution of events to the N_T timers is done deterministically. Therefore, the relative variance in the number of photons per timer is smaller than that for the detectors ($\sqrt{N_T}$ and N_D, resp.), making a higher throughput feasible.

Re-running the simulations to take the combined effects of slow detectors and timing pile-up into account (Arch. IV without router pile-up) leads to virtually identical results for the number of SPADs and timers needed (also shown in Fig. 6). This also includes a single detector coupled to a single timer as a special case, where it is found that the CMM lifetime estimate stays within 1% error margins up to a photon rate of $\mu = 4.5\%$. To be able to measure higher photon rates using this classical architecture, N_C of such channels have to be used in parallel (Arch. I), where N_C has to increase linearly with the photon rate. This requires a total of $N_D = N_C$ detector

elements and $N_T = N_C$ timing elements. In contrast, if a sensor architecture with "routed" timers is used, a much smaller number of timers is sufficient to achieve the same throughput.

D. Router pile-up

However, as soon as events have to be routed between different detectors and/or timers, additional distortions will be introduced. Pulses from multiple detection elements get shortened but still have a finite length t_P and therefore can merge into a single pulse when they get combined in the router. If two photon events were registered less than t_P apart, the combined pulses merge together and only the first event will get processed further and the second event is lost completely. For a single exponential decay, the probability for router pile-up can be calculated analytically. As long as "timer pile-up" is negligible, the detection probability can be expressed in closed form as

$$c_{\mathrm{tot}}(\mu;t) = \frac{1}{\tau}\exp(-t/\tau) \times \begin{cases} \exp(-\mu(1-e^{-t/\tau})) & t < t_P \\ \exp(-\mu e^{-t/\tau}(e^{+t_P/\tau}-1)) & t \geq t_P \end{cases} \quad (2)$$

(for details on derivation, see the Appendix).

Initially, at most one photon can be detected and therefore the probability is exactly the same as for a classical TCSPC system (see Eq. (7.20) in Ref. 1), but after a time delay corresponding to the shortened SiPM pulse width t_P additional photons can be detected. The resulting histogram is evidently no longer a single exponential decay, but after the initial distortion it quickly recovers to approximate the true decay well. This is illustrated in Figure 8, which shows data from the MATLAB model in very good agreement with the analytical result.

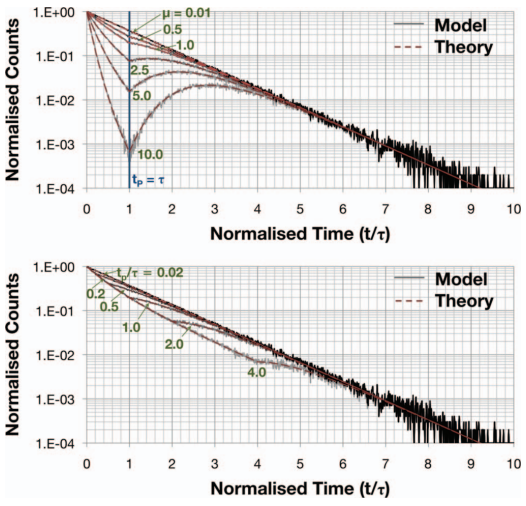


FIG. 8. Effect on the captured histogram of varying μ for fixed $t_P = \tau$ (top) and t_P/τ for fixed $\mu = 1.0$ (bottom), using the model (solid) and theory (dashed).

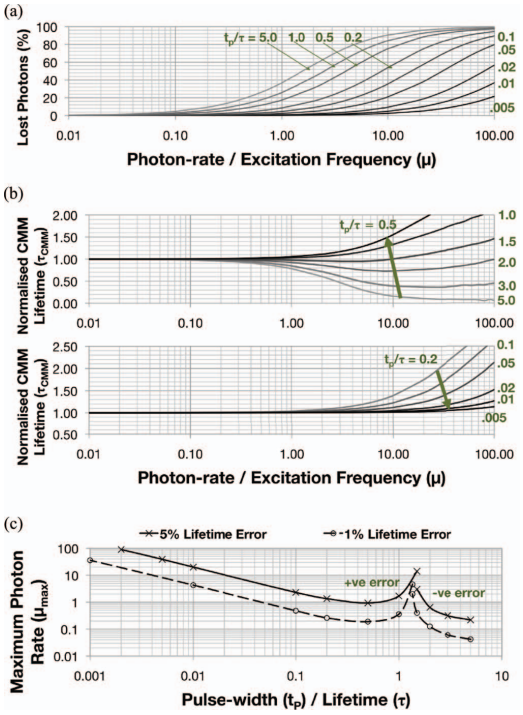


FIG. 9. (a) Percentage of lost photons and (b) CMM lifetime estimates for a sensor showing router pile-up for different ratios of pulse width t_P to the measured lifetime. (c) Maximal achievable throughput keeping the lifetime estimation error below 1% or 5% as a function of t_P/τ.

The number of detected counts per laser cycle can then be calculated from the above as

$$c_{\mathrm{int}}(\mu) = \mu\int_0^\infty c_{\mathrm{tot}}(\mu;t')dt' = \frac{e^{t_P/\tau}}{e^{t_P/\tau}-1}[1-\exp(-\mu(1-e^{-t_P/\tau}))]. \quad (3)$$

This expression, also illustrated in Figure 9(a), clearly shows that photon loss caused by router pile-up is almost negligible in the case of t_P small compared to the lifetime τ, whereas for very large t_P the performance deteriorates to that of a classical TCSPC system with a maximum of 1 photon per cycle detected. For large t_p, the lifetime is under-estimated just as in the classical case, but if t_p becomes smaller than the measured lifetime τ the estimates become more accurate even at very high count rates (Fig. 9(b)).

In a real sensor, the pulse width t_P will be fixed by the sensor design, so the performance of the sensor will change with the lifetime τ that is being measured. The throughput with a routed channel always exceeds that of a traditional TCSPC system (i.e., $\mu_{1\%} = 0.045 f_E$) but if the lifetime is short compared to t_P, the performance increase is fairly small (Figure 9(c)). However, if the lifetime is larger than about $5–10t_P$, the systematic errors at all realistically achievable count rates become negligible. For best performance, the

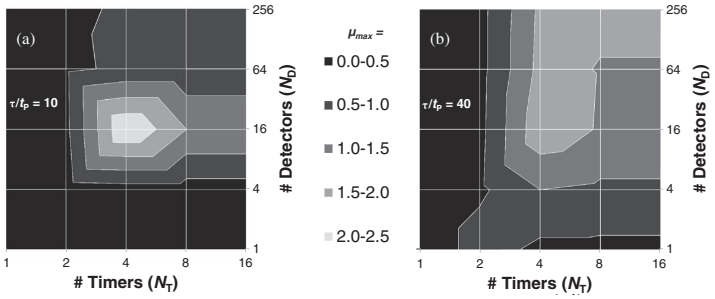


FIG. 10. Maximum achievable photon-rate ($\mu_{\max}$) for a 1% lifetime calculation error as a function of N_T and N_D for a detector with $t_D/t_P = 40$. (a) Short ($\tau = 10t_P$) and (b) long lifetime ($\tau = 40t_P$). Note that the throughput always exceeds that of parallel sensors (Arch. I), where $\mu = 0.045N$.

pulse width t_P has to be kept as short as possible if relatively short lifetime decays need to be measured.

E. Combined effects

In a real chip, all the different forms of pile-up will occur simultaneously. In general, there is a non-trivial interplay between these distortions, which now not only depend on the number of detector elements and timers but also on the lifetime that is being measured.

Full exploration of these combined effects is beyond the scope of this paper, but to illustrate the performance for realistic experimental parameters, two contour plots of the maximum achievable count rates for a given lifetime estimation error of 1% are shown in Figure 10. If the measured lifetime is not much larger than t_P, the maximum throughput does not increase monotonically with the number of timers and/or detectors as systematic errors of the various pile-up mechanisms cancel each other out. Although this can lead to a sharp increase in throughput (local maximum in Fig. 10(a)), the performance depends rather sensitively on the measured lifetime τ. However, if t_P/τ is small ($\lesssim 2\%$), the router pile-up becomes negligible and the performance of the detector becomes independent of the measured lifetime and can actually reach the throughput of the ideal integrated detector as shown in Figure 6.

Thus, an integrated sensor using a "routed" architecture (Arch. IV) performs best for long lifetimes (compared to t_P), which is where pile-up problems are encountered most frequently in the first place. For short lifetimes, there is less benefit of such a sensor, especially as in this case pile-up can be easily minimised by increasing the excitation frequency.[5,6]

IV. EXPERIMENTAL RESULTS

A. The sensor

Details of the sensor hardware implementation have been presented elsewhere,[9] so its characteristics will only be summarized briefly. The sensor comprises a 1.3 × 1.7 mm CMOS chip in 0.13 μm technology, integrating an array of SPADs arranged as a SiPM, a multiple channel time-to-digital converter architecture, and embedded CMM pre-processing of the high bandwidth data (see Fig. 3).

The SiPM contains 1024 (32 × 32) 8 μm diameter active area SPAD devices[14] with a 21.5 μm pitch, providing a total active area of over 0.5 mm^2 with over 10% fill factor. Individual SPADs can be enabled or disabled independently allowing a suitable group of low DCR detectors to be selected for experimentation. To reduce the effect of SPAD dead-time ($t_D \approx 10$ ns), the detector pulses are shortened to a width $t_P \approx 540$ ps and then passed through a balanced OR-tree to provide a single sensor output to the time-resolving circuitry. Time-resolved measurements are performed using an array of 16 TDCs with ≈52 ps resolution[15] and an extended range of up to 3.6 μs (16 bits). To eliminate the effects of timing dead-time t_T, one half of the TDCs are available on alternate excitation periods, making it possible to time up to eight photon events per excitation period.

Due to the high data rates of over 1200 Mbps (with a 10 MHz laser) produced by the TDCs, embedded on-chip processing is required. Therefore, a pre-calculation of the centre-of-mass is integrated on chip. As can be seen from the expression introduced earlier (Eq. (1)), the algorithm requires a summation of the TCSPC codes and a count of the valid TCSPC events N_c in the decay histogram within a pre-defined window of duration T. The final division and background correction is performed externally using a micro-controller, FPGA or in software depending on application requirements.

The device can be configured to provide full TDC codes for up to 1 event per laser cycle, CMM data for up to 1 event per laser cycle, or CMM data for up to 8 events per laser cycle. Results from raw TDC code capture can be found in Ref. 9 and the other two capture modes will be used as a comparison of the pile-up resistance in this paper.

B. Experimental setup

For experimental evaluation of the sensor's performance, we measured a range of fluorescent dyes for varying laser excitation power. The sensor acts as a point detector but for experimental convenience the measurements were performed in an existing fluorescence lifetime wide field imaging system on a Nikon TE2000U inverted microscope. The excitation source is a PicoQuant pulsed diode laser with a wavelength of

478 nm coupled through the epi-fluorescence port of the microscope using a Nikon B-2A filter cube. The laser pulse repetition rate is $f_E = 10$ MHz and the maximum power reaching the back focal plane of the objective is about 64 μW. The excited sample volume is focussed onto the SiPM detector attached to one of the camera ports using an additional short focal length lens. As each individual SPAD contributes dark counts, the noise floor increases with the number of active detector. So rather than having all detectors active the light is concentrated on a small group of SPADs such that each individual SPAD obtains a similar count rate. Experimental results presented in the following used a group of 8 SPADs with a combined dark count rate of 5.5 kHz. The TDC range was kept fixed at 115.5 ns (time bin width of $R_T = 0.113$ ns) for all the measurements. The incident laser power is changed by combining various neutral density filters in the excitation path.

C. Experimental results

As discussed in Sec. III D, the sensor's performance will depend on the fluorescence lifetime (compared to sensor's shortened pulse duration t_P), so samples with different lifetimes were examined. Photon detection rates and CMM based lifetime estimates were recorded as a function of incident light for four samples covering most of the routinely used range of lifetimes (see Table II). The fluorescence light incident on the detector was varied by adjusting the intensity of the excitation laser light. At low excitation intensities, the detector count rate increases linearly with laser power, which is used to estimate the number of "detectable" photons r as a function of laser power for each given sample (Table II). For the lifetime estimation, the measurement time window is defined by the indices of its first and last time bin, FIRST and LAST. The parameter LAST was fixed at 2 time bins below the peak position of the decay, whereas FIRST was adjusted for the different fluorophores such that the duration of the time window $T = (\text{LAST} - \text{FIRST}) * R_T$ remained sufficiently large compared to the lifetime τ. As discussed in detail in previous publications,[12,16] for small ratios of T/τ additional correction factors are required to ensure the accuracy of the lifetime estimate. On the other hand, if the window T is too large, the signal to noise ratio decreases. The CMM lifetime estimates corrected for background counts and finite time window T agree well with accepted literature values[17] and single exponential tail fits to histogram data acquired with our sensor.

TABLE II. Details of the experimental parameters for the four different samples used for testing the sensor. Quoted CMM lifetime estimates are based on multi-channel results at about 1 kHz count rate, 1 ms exposure time, and 1 s total acquisition time, corrected for the finite size of time window T. The estimated number of photons per excitation power r is also shown.

Fluorophore/solvent	τ [ns]	T [ns]	r [1/μW]
Rhodamine B/water	1.79 ± 0.08	9.5	1050
Rhodamine 6G/water	3.96 ± 0.14	33.3	1060
Rubrene/methanol	8.56 ± 0.25	39.9	370
QDots/toluene	16.5 ± 0.5	44.6	1700

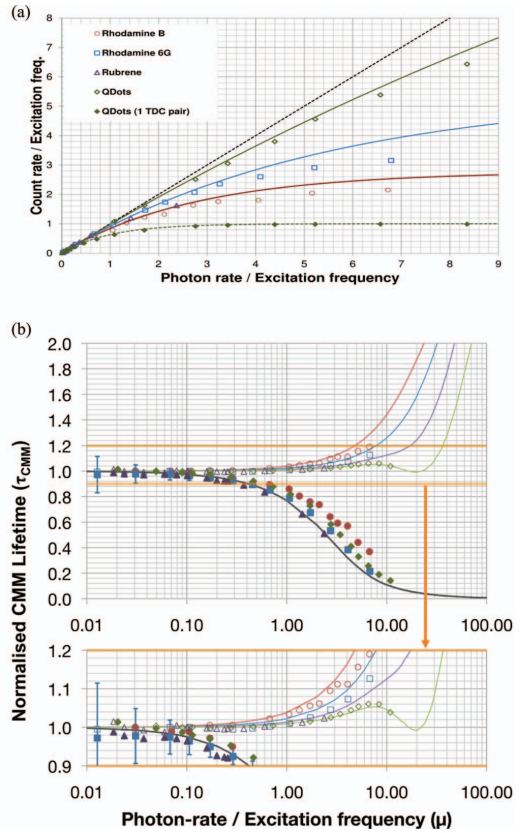


FIG. 11. (a) Recorded count rates as a fraction of "detectable" incident photons for fluorophores of different lifetimes (Rhodamine B (red circles), Rhodamine 6G (blue squares), Rubrene (purple triangles), and quantum dots (green diamonds)). (b) Normalised background-corrected fluorescence lifetime estimates for the case of single-channel (solid symbols) or multi-channel (open symbols) timing mode together with simulated results (lines).

Figure 11 shows the combined experimental results. Lifetime estimation was performed in either traditional single channel mode (only one TDC pair active) or multi-channel mode with all 8 TDC pairs enabled. Count rates in single-channel mode are determined from the number of timed events, whereas for the multi-channel mode the timing-circuits were bypassed and the merged stream of shortened SPAD pulses was directly counted. All lifetime estimates are based on 100 μs exposure intervals, with mean values and standard deviation determined from 10 000 samples, i.e., a total acquisition time of 1 s.

The normalised CMM lifetime results, shown by the unfilled markers in Figure 11, clearly demonstrate the sensor's ability to more accurately calculate the lifetime value at high photon-throughputs. In all cases, a photon throughput equal to the excitation rate is demonstrated for a worst case error of 4% for the shortest lifetime (Rhodamine B) and a best case of only 1% for the longer lifetime fluorophores. Furthermore, a photon throughput of five times the excitation frequency is

possible for the quantum dot sample for a 5% error in calculation, where the single channel could only achieve about $0.3f_E$ for the same error. The figure also shows numerical results of the MATLAB model based on the experimental parameters (see Table I) as well as the actual fluorescence lifetimes. The similarities between the model and laboratory results are clearly shown at the bottom of Figure 11 by the solid coloured curves.

V. CONCLUSIONS

We analyzed the performance of several integrated fluorescence lifetime sensor designs using a MATLAB model. This demonstrated that a router based design can out-perform traditional designs and potentially achieve very high photon throughput. The experimental performance of our novel sensor matches modeled data and analytical expressions, and we showed that it can produce reliable lifetime estimates at photon count rates well beyond the classical pile-up limit. Our sensor can therefore provide faster lifetime measurements and operate with a much higher dynamic range than classical TCSPC detectors. It should therefore be well suited for applications that require high speed and a good dynamic range such as confocal (or multiphoton) scanning fluorescence lifetime imaging. It should also enable flow cytometry or cell sorting based on fluorescence lifetime information and dramatically accelerate TCSPC data acquisition for long lifetime dyes such as ruthenium based oxygen sensing fluorophores.[18]

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (U.K.) EPSRC(GB). The authors would like to thank STMicroelectronics for chip fabrication. The assistance of Richard Walker and Abigail Johnston's help in testing an early prototype of the sensor is gratefully acknowledged.

APPENDIX: DERIVATION OF THEORETICAL EXPRESSIONS

Emissions of fluorescence photons can be described by a Poisson distribution. If μ is the average number of photons per laser cycle incident on the detector, then

$$f_\mu(n) = \frac{\mu^n}{n!}\exp(-\mu)$$

describes the probability of having n photons arriving within any given laser cycle.

While the probability of having multiple photons arriving within a signal cycle is negligible for very low μ, it generally has to be taken into account as additional photons in a given laser cycle lead to photon loss. For an ideal sensor, the probability of detecting a photon which arrives as the mth out of a total of n photons is given by

$$P_{m,n}(t) = w_{m,n} \times [P_{\text{before}}(t)]^{m-1} \times P(t) \times [P_{\text{after}}(t)]^{n-m} \tag{A1}$$

with $P_{\text{before}}(t)$, $P(t)$, $P_{\text{after}}(t)$ the probability for a photon to arrive before, at, or after delay time t, respectively, and $w_{m,n}$ the statistical weight. As the photons are indistinguishable, the weight is simply given by a binomial coefficient $w_{m,n} = \binom{n-1}{m-1}$.

However, for a real sensor the system is blind for a dead time t_P, so if any photon has arrived within the time interval from $\max(0, t - t_P)$ to t, the mth photon will no longer get detected. The *mth* photon will therefore only be detected with a reduced probability which can be expressed by replacing $P_{\text{before}}(t)$ with $P_{\text{before}}(t - t_P)$ (which is assumed to be 0 for $t < t_P$)

$$P_{\text{det}\,m,n}(t) = w_{m,n} \times [P_{\text{before}}(t - t_P)]^{m-1} \times P(t) \times [P_{\text{after}}(t)]^{n-m}. \tag{A2}$$

So the total number of detected photon events due to the mth arriving photon within a signal cycle is

$$c_m(\mu; t) = \sum_{n=m}^{\infty} n f_n(\mu) P_{\text{det}\,m,n}(t) \tag{A3}$$

and the total number for all photons within a signal cycle is

$$c_{tot}(\mu; t) = \sum_{n=m}^{n_{\max}} c_m(\mu; t), \tag{A4}$$

where $n_{\max}$ is the maximum number of photons that can be timed within a signal cycle.

For a single exponential decay $P(t) = 1/\tau \exp(-t/\tau)$, it is fairly straightforward to express most of these expressions analytically. With

$$P_{\text{before}}(t - t_P) = \int_0^{t-t_P} P(t')dt' = \begin{cases} 0 & t < t_P \\ 1 - \exp(-(t - t_P)/\tau)) & t \geq t_P, \end{cases}$$

$$P_{\text{after}}(t) = \int_t^{\infty} P(t')dt' = \exp(-t/\tau),$$

one finds for the count rate due to the photons which arrive first at the detector

$$c_1(\mu; t) = \frac{1}{\tau} e^{-t/\tau} \sum_{n=0}^{\infty} e^{-nt/\tau} f_\mu(n) = \frac{1}{\tau} e^{-t/\tau} \exp(-\mu(1 - e^{-t/\tau})) \tag{A5}$$

and for a later, mth, photon

$$c_m(\mu; t) = \frac{1}{\tau} e^{-t/\tau} \sum_{n=m}^{\infty} n \binom{n-1}{m-1} (1 - e^{-(t-t_P)/\tau})^{m-1} \times e^{-(n-m)t/\tau} f_\mu(n) \text{ for } t \geq t_P, \tag{A6}$$

This can also be expressed as a recursive relation for $m > 1, t \geq t_p$:

$$c_m(\mu; t) = \frac{1}{\tau} \frac{\mu}{m-1} (1 - e^{-t/\tau} e^{t_P/\tau}) c_{m-1}(\mu; t). \tag{A7}$$

If all photons can be timed ($m_{\max} = \infty$), the total count rate can also be expressed analytically

$$c_{\text{tot}}(\mu; t) = \frac{1}{\tau} \exp(-t/\tau) \times \begin{cases} \exp(-\mu(1 - e^{-t/\tau})) & t < t_P \\ \exp(-\mu e^{-t/\tau}(e^{+t_P/\tau} - 1)) & t \geq t_P \end{cases}. \quad \text{(A8)}$$

For $t < t_P$, only one photon can be detected and therefore the probability is exactly the same as for a classical TCSPC system (see Eq. (7.20) in Ref. 1), but after the shortened detector dead time t_P additional photons can be detected.

The average number of counts per laser cycle can then be calculated as

$$c_{\text{int}}(\mu) = \mu \int_0^\infty c_{\text{tot}}(\mu; t')dt' = \frac{e^{t_P/\tau}}{e^{t_P/\tau} - 1}[1 - \exp(-\mu(1 - e^{-t_P/\tau}))]. \quad \text{(A9)}$$

[1] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*, Springer Series in Chemical Physics Vol. 81 (Springer Verlag, New York, 2005).
[2] W. Becker, B. Su, and A. Bergmann, Proc. SPIE **7183**, 718305 (2009).
[3] P. Coates, J. Phys. E **1**, 878 (1968).
[4] V. Katsoulidou, A. Bergmann, and W. Becker, Proc. SPIE **6771**, 67710B (2007).
[5] L. Turgeman and D. Fixler, "The influence of dead time related distortions on live cell fluorescence lifetime imaging (FLIM) experiments," J. Biophotonics (published online).
[6] D. McLoskey, D. Campbell, A. Allison, and G. Hungerford, Meas. Sci. Technol. **22**, 067001 (2011).
[7] C. Davis and T. King, J. Phys. A **3**, 101 (1970).
[8] O. M. Williams and W. J. Sandle, J. Phys. E **3**, 741 (1970).
[9] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, in *Proceedings of the 2012 IEEE International Solid-State Circuits Conference* (IEEE, 2012), pp. 122–124.
[10] R. A. Colyer, G. Scalia, F. A. Villa, F. Guerrieri, S. Tisa, F. Zappa, S. Cova, S. Weiss, and X. Michalet, Proc. SPIE **7905**, 790503 (2011).
[11] C. Veerappan, J. Richardson, R. Walker, D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, in *Proceedings of the 2011 IEEE International Solid-State Circuits Conference* (IEEE, 2011), pp. 312–314.
[12] D.-U. Li, B. Rae, R. Andrews, J. Arlt, and R. Henderson, J. Biomed. Opt. **15**, 017006 (2010).
[13] D. Tyndall, "A CMOS system for high throughput fluorescence lifetime sensing using time correlated single photon counting," Ph.D. thesis, The University of Edinburgh, 2013.
[14] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, IEEE Trans. Electron Devices **58**, 2028 (2011).
[15] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, in *Proceedings of the Custom Integrated Circuits Conference (CICC)* (IEEE, 2009), pp. 77–80.
[16] D. D.-U. Li, J. Arlt, D. Tyndall, R. Walker, J. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, J. Biomed. Opt. **16**, 096012 (2011).
[17] N. Boens, W. Qin, N. Basarić, J. Hofkens, M. Ameloot, J. Pouget, J.-P. Lefèvre, B. Valeur, E. Gratton, M. VandeVen, N. D. Silva, Y. Engelborghs, K. Willaert, A. Sillen, G. Rumbles, D. Phillips, A. J. W. G. Visser, A. van Hoek, J. R. Lakowicz, H. Malak, I. Gryczynski, A. G. Szabo, D. T. Krajcarski, N. Tamai, and A. Miura, Anal. Chem. **79**, 2137 (2007).
[18] H. C. Gerritsen, R. Sanders, A. Draaijer, C. Ince, and Y. K. Levine, J. Fluoresc. **7**, 11 (1997).

B.2 Tyndall et. al., T. Bio. CAS, 2012

A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13 μm CMOS

David Tyndall, *Student Member, IEEE*, Bruce R. Rae, *Member, IEEE*, David Day-Uei Li, Jochen Arlt, Abigail Johnston, Justin A. Richardson, *Member, IEEE*, and Robert K. Henderson, *Member, IEEE*

***Abstract*—We describe a miniaturized, high-throughput, time-resolved fluorescence lifetime sensor implemented in a 0.13 μm CMOS process, combining single photon detection, multiple channel timing and embedded pre-processing of fluorescence lifetime estimations on a single device. Detection is achieved using an array of single photon avalanche diodes (SPADs) arranged in a digital silicon photomultiplier (SiPM) architecture with 400 ps output pulses and a 10% fill-factor. An array of time-to-digital converters (TDCs) with $\approx$ 50 ps resolution records up to 8 photon events during each excitation period. Data from the TDC array is then processed using a centre-of-mass method (CMM) pre-calculation to produce fluorescence lifetime estimations in *real-time*. The sensor is believed to be the first reported implementation of embedded fluorescence lifetime estimation. The system is demonstrated in a practical laboratory environment with measurements of a variety of fluorescent dyes with different single exponential lifetimes, successfully showing the sensor's ability to overcome the classic *pile-up* limitation of time-correlated single photon counting (TCSPC) by over an order of magnitude.**

***Index Terms*—Biosensors, silicon photomultipliers (SiPMs), single photon avalanche diodes (SPADs), time-correlated single photon counting (TCSPC), time domain fluorescence lifetime.**

I. Introduction

INTEGRATION of fluorescence lifetime sensing on silicon is enabling advances in cell-biology research, medical diagnosis and pharmacological development [1]. As well as being independent of probe concentration, illumination intensity and emission wavelength, fluorescence lifetime can be used to acquire more quantitative information about physiological parameters such as pH levels [2] and Ca^{2+} concentrations [3].

Time-correlated single photon counting (TCSPC) [4], [5] is the most photon-efficient technique for measuring fluorescence lifetime in the time-domain [6]. It is performed by repeatedly timing fluorescence emission with respect to a synchronized pulsed optical excitation, to build a histogram of the lifetime decay. However, a major limitation of this approach is the restrictively low photon throughput limit of $\approx$ 5% of the excitation rate [7], which is necessary in order to avoid distortion of the decay histogram due to *pile-up* [8]. Photon throughput rates of several MHz have been achieved experimentally by using a high repetition rate (100 MHz) excitation source [9], however this approach is restricted to measuring fluorescent dyes with lifetimes below 2 ns.

Fluorescence lifetime sensing can also be performed in the frequency-domain, by measuring amplitude and phase changes between sinusoidal optical excitation and detected sinusoidal emission [10]. This technique is capable of higher photon throughput, necessary for applications such as flow cytometry [11] and fluorescence lifetime imaging (FLIM) of bright samples [12]. However, hardware approaches to high-throughput fluorescence lifetime sensing in the time-domain, which is better suited to an integrated, low power implementation, have not been fully explored. This paper introduces techniques to overcome *pile-up* in the time-domain, enabling an increase in available throughput for applications such as flow cytometry and FLIM as well as high throughput screening (HTS) [13] and functional near-infrared spectroscopy (fNIRS) [14].

Typical TCSPC apparatus includes a pulsed optical source, a discrete detector such as a photomultiplier tube (PMT) or silicon photomultiplier (SiPM) [15], external time-to-digital conversion (TDC) and histogramming hardware, and a CPU to compute the decay constant. Multimodule systems, consisting of many discrete detectors and an equal number of independent timing modules, provides one solution to the *pile-up* problem. However, such setups significantly increase the system size, cost, complexity and power consumption [4].

Recent advances in single-photon avalanche diodes (SPADs) and TDCs manufactured in standard CMOS processes have enabled TCSPC measurements to be performed by an imaging array [16]; however this device produces data at over 25 Gb/s and has a low fill factor of $<$2%. System miniaturization has been achieved in CMOS using time-gated fluorescence lifetime

Manuscript received May 18, 2012; revised August 20, 2012; accepted September 30, 2012. Date of publication December 20, 2012; date of current version January 14, 2013. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) and STMicroelecronics Imaging Division, Edinburgh. This paper was recommended by Associate Editor M. Ghovanloo.

D. Tyndall is with the CMOS Sensors and Systems Group and the Institute for Integrated Micro and Nano Systems, School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, U.K., and also with Dialog Semiconductor, Edinburgh EH1 3DQ, U.K (e-mail: d.tyndall@ed.ac.uk).

B. R. Rae is with STMicroelectronics, Pinkhill, Edinburgh EH12 7BF, U.K.

D. D.-U. Li is with the Department of Engineering and Design, School of Engineering and Informatics, University of Sussex, Brighton BN1 9QT, U.K.

J. Arlt and A. Johnston are with COSMIC, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3JZ, U.K.

J. A. Richardson is with Dialog Semiconductor, Edinburgh EH1 3DQ, U.K.

R. K. Henderson is with the CMOS Sensors and Systems Group and the Institute for Integrated Micro and Nano Systems, School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBCAS.2012.2222639

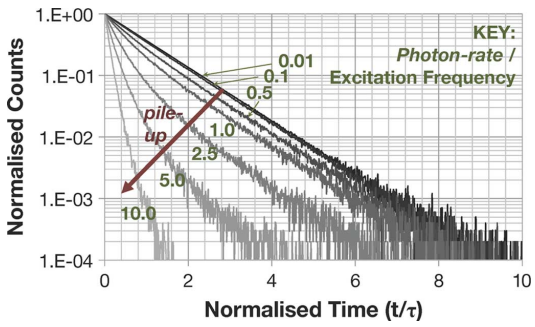


Fig. 1. Simulated effect of *pile-up* on TCSPC histograms.

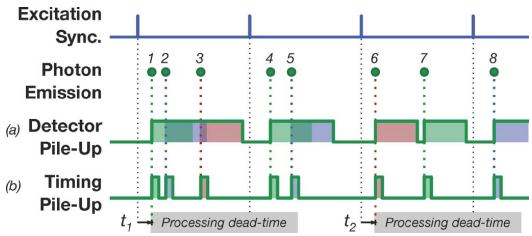


Fig. 2. TCSPC Detector and Timing *pile-up*.

sensing techniques to reduce data bandwidth and processing requirements [17], [18], however such approaches are less photon-efficient than TCSPC and are still limited by *pile-up* due to detector dead-time.

The causes and effects of TCSPC *pile-up*, together with techniques to overcome them, are introduced in Section II. Implementation details of the fabricated device, designed to overcome TCSPC *pile-up*, are detailed in Section III, before non-idealities with the timing architecture are discussed in Section IV. Finally, an overview of the experimental setup and details of results are given in Sections V and VI, showing the device's ability to achieve photon throughputs exceeding the excitation rate, whilst allowing lifetimes of common fluorescent markers to be obtained.

II. TCSPC *Pile-Up*

The TCSPC *pile-up* phenomenon is caused by a combination of long detector dead-times and the use of only one timing element in the processing hardware [4]. As the detected *photon-rate* is increased towards and past the excitation rate, the hardware begins to miss events that occur late in the decay, distorting the histogram towards shorter times. This is shown by the simulated TCSPC histogram data in Fig. 1. The effect of *pile-up* is minimized by keeping the *photon-rate* <0.05 of the excitation rate. The term *photon-rate* is used to refer to the number of photons that cause the active area of the detector to register an event and is directly proportional to the emission intensity.

Detector *pile-up* occurs at high *photon-rates* when there is an increased probability of a photon event arriving during the dead-time of the previous one, making them appear to overlap at the detector output. This is shown conceptually in Fig. 2(a), where the positive edges of photon events 2, 3 and 5 are lost within the dead-time of the event that preceded them. For passively quenched SPAD detectors, this dead-time is in the region of tens of nanoseconds [19].

By using pulse-shortening circuitry at the output of the detector, sub-nanosecond pulses can be generated [20]. In this case, the single timing element becomes the major limiting factor. As shown by the conceptual example in Fig. 2(b), the single timing element and its processing dead-time (shown by the shaded region) only allow photon events 1 and 6 to be recorded, producing time-stamps t_1 and t_2. The processing dead-time is in the region of hundreds of nanoseconds to microseconds [4]. Combining sub-nanosecond detector output pulses with multiple timing elements to process many photon events concurrently, as proposed by [5], will significantly increase the possible throughput of TCSPC.

Although sub-nanosecond detector output pulse-widths improve the photon throughput considerably, there is still a likelihood of events overlapping at higher *photon-rates*. For a fixed pulse-width, the chances of overlap are heavily dependent on the lifetime of the fluorophore being measured, with longer lifetime fluorophores being affected less than short. Therefore, this approach is more suited to measuring longer lifetime fluorophores, which at present are most limited by classic TCSPC *pile-up* due to the extended excitation periods required.

Pile-up distortion can be corrected post-experiment in software [21], however, these techniques require additional hardware to monitor the total photon count rate in parallel with acquiring the TCSPC data, including events that were not processed by the timing hardware [22]. Furthermore, as well as this technique requiring additional hardware and processing, it remains limited by detector *pile-up* at higher *photon-rates*.

III. Device Implementation

A. Overview

The fully digital single-chip solution to time-resolved fluorescence lifetime sensing [23], comprises a 1.3×1.7 mm^2 CMOS chip in 0.13 μm technology, as shown in Fig. 3. The device features a digital SiPM [24] with shortened output pulses, a multi-channel TDC architecture and embedded single exponential lifetime estimation processing using a centre-of-mass method (CMM) [25] pre-calculation. The proposed architecture, as shown in Fig. 4, is designed to allow the *pile-up* limit to be increased by over an order of magnitude.

B. Silicon Photomultiplier

A digital mini-SiPM architecture with pulse-shortened output is employed to perform the single photon detection. The SiPM comprises 1024 circular, 8 μm diameter active area, negatively biased, passively quenched SPADs from [19], arranged in a 32×32 array. The large number of detectors in the SiPM was chosen to provide experimental flexibility, with only a small subset expected to be used for experiments in order to reach a target throughput of 100 MHz. Detector sensitivity is controlled using an external high-voltage excess bias, V_{EB} (2.8–3.3 V), which powers the SPAD, requiring thick-oxide transistors

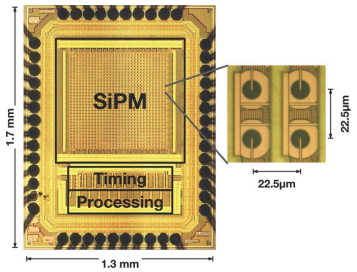


Fig. 3. Annotated micrograph of fabricated chip showing 4 SPAD detectors.

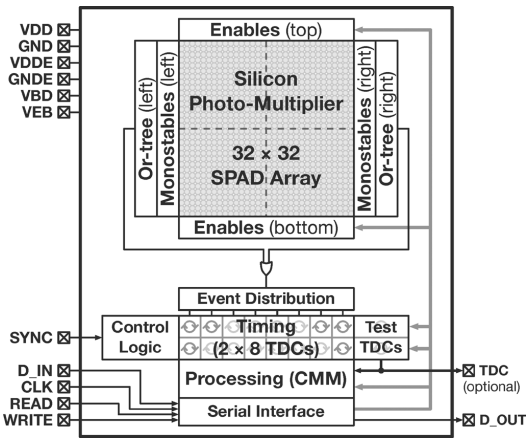


Fig. 4. Device Top-Level Block Diagram.

within the array. This external power supply is controlled independently of the core V_{DD} (1.2 V) supply, protecting the timing and processing circuitry from light dependent power consumption.

To implement the pulse-shortening, which prevents the dead-time of individual detectors in the SiPM from restricting the maximum count rate, the buffered output from each individual SPAD detector is passed through a monostable circuit, as shown both as a concept (top) and as implementation (bottom) in Fig. 5. The monostable circuit is implemented by NORing (I5) the SPAD pulse with a delayed version of itself (delay through $\mathrm{I2} + \mathrm{I3} + \mathrm{I4}$), where the pulse width at the output is equal to the total delay time.

The monostable circuit has an inverter with thick-oxide transistors (I1) powered by V_{DD}, whose input is overdriven to level shift the SPAD output back to the core supply voltage. Additionally, all of the pulse-shortening circuits can be enabled or disabled using the global $\overline{\mathrm{MS\_EN}}$ signal. Although the output of the monostable circuits will have short pulse widths, the individual SPADs that created these events will remain insensitive to subsequent photon events within their own inherent dead-time. This creates a spatial *pile-up* which is minimized by the use of several small active area detectors within the SiPM architecture. This ensures that multiple photon events occurring

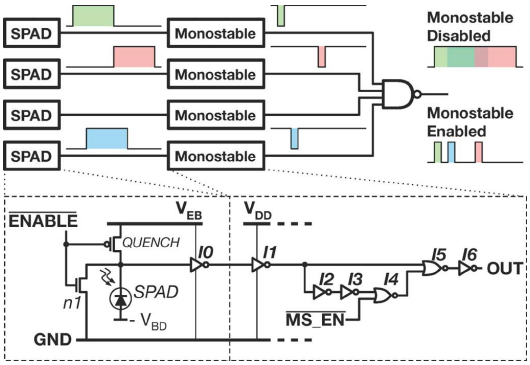


Fig. 5. Monostable pulse-shortening concept with circuit implementation.

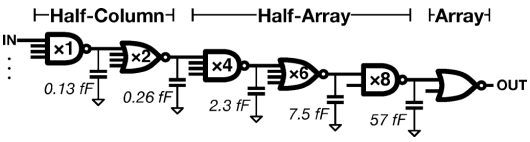


Fig. 6. OR-tree circuit, showing increasing buffer strengths.

within a single excitation period are more likely to fall on different detectors within the SiPM [4].

Each shortened SPAD pulse is then sent through a balanced logical OR-tree to provide a single detector output, as shown in Fig. 6. To minimize delay for the timing critical TCSPC application, the OR-tree is implemented in negative logic using *High-Speed* standard cells with increasing buffer strength to compensate for the increasing length of track—and therefore increasing capacitance—that each stage must drive.

The increasing track length also limits the monostable pulse-width as it must be long enough to safely pass through the tree and trigger the timing circuitry at the output. After thorough extracted SPICE simulations of the worst-case pulse path through the OR-tree, a monostable output pulse-width of ≈ 400 ps was chosen. This pulse-width is created by increasing the minimum transistor gate width and length of the NMOS and PMOS in the delay inverter (I3) in Fig. 5.

Each SPAD in the SiPM can be individually and independently turned on or off using static enable signals. This not only provides a method for monitoring each SPAD individually to measure performance characteristics such as dark count rate (DCR) and timing, but also acts as a method to enable a sub-array for experimentation. Noisy detectors that would otherwise have a negative effect on the SNR can also be disabled, at the expense of reduced sensitivity.

The SiPM is partitioned with the SPAD, quench and output buffer, as shown to the bottom left of Fig. 5, being located inside the array. The monostable, OR-tree and enable circuits are then located at the periphery of the array as shown in Fig. 4. This partitioning was chosen to maximize the fill-factor of the detector whilst keeping circuitry that is critical to the timing performance local to the SPAD. The timing critical components include the output buffer, which must drive long, high capacitance tracks

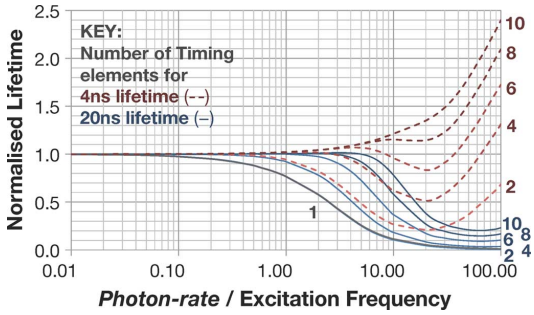


Fig. 7. Effect of increasing the *photon-rate* on CMM calculation.

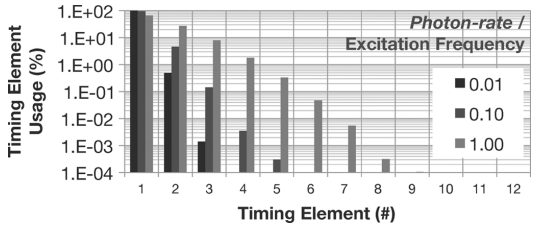


Fig. 8. TDC usage for increasing *photon-rates*.

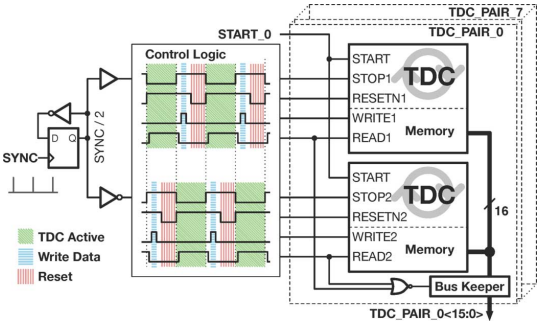


Fig. 9. Interleaved TDC-pair timing.

and the SPAD quenching element which controls the detector dead-time. The resulting pitch of the SPADs and corresponding circuitry in the array is 21.5 μm, as shown in Fig. 3. Within the large SiPM area, it is possible to locate a small subset of adjacent low DCR SPADs (e.g., a 3×3 grid) required for experimentation and therefore provide a fill-factor of just over 10%.

C. Multiple-Channel Timing

In order to process multiple events from the pulse-shortened output, high throughput SiPM within a single excitation period, a multiple-channel timing architecture is required. To understand the effect that the number of timing channels in the architecture has on the extracted lifetime calculation, a detailed model of the device is created. This time-domain MATLAB model produces histogram and CMM data based on photon count-rate, SiPM pulse-width, fluorescence lifetime, excitation rate, number of TDCs and TDC resolution, whilst also looking at *pile-up* losses and TDC usage. The model is run with 4 ns and 20 ns lifetimes, a 20 MHz excitation rate and 1 000, 50 ps resolution time bins. The SiPM output pulse is set at 400 ps, as discussed in Section III.B.

The graph in Fig. 7 shows the effect of increasing the *photon-rate*, for a varying number of timing elements, on the normalized CMM calculation. The range of *photon-rates* over which the lifetime can be correctly calculated is shown to increase by adding timing elements. However, there are clear differences between the lifetime estimations for 4 ns (dashed) and 20 ns (solid) lifetimes. At high *photon-rates*, the fixed pulse-width causes detector *pile-up* to be the limiting factor for the 4 ns lifetime. However, the lack of timing elements is the limiting factor for the 20 ns lifetime, where events are more spread-out.

Fig. 8 shows the percentage of photons processed by each timing element for a 20 ns lifetime with varying *photon-rates*. Although four timing elements allows $\approx$ 99% of photons to be processed for a *photon-rate* equal to the excitation repetition rate, it can be seen that eight timing elements will allow almost *every* photon (99.999%) to be processed and will allow the throughput to exceed the excitation repetition rate.

The timing element of this architecture is a ≈ 50 ps resolution time-to-digital converter (TDC) from [26], chosen for its small area footprint and ability to provide timing information in *real-time* with no latency. It has been modified to extend the full range from 10 to 16 bits, allowing lifetime decays to span over 3 μs. This allows much longer lifetimes to be measured, taking advantage of the suitability of the chosen architecture to speed up the acquisition of TCSPC for long lifetime fluorophores. The TDC will operate in a reversed start-stop mode, so that power consumption is proportional to photon activity [4].

Similar to typical TCSPC timing hardware, the TDC requires a finite processing dead-time to save its data and be reset ready to receive the next photon event. To remove this dead-time, the TDCs are arranged in pairs with interleaved timing so that one TDC is active and available to accept an event while the other is writing its data and being reset. The TDCs are entirely controlled by the excitation synchronization pulse delivered either by the excitation source or by FPGA. As the pulse typically provided by a laser is a short impulse, it is initially divided by two using a toggle flip-flop, as shown in Fig. 9, creating two out of phase clocks, whose positive going edges act as the *STOP* signal to the pair of TDCs. Each TDC is therefore only active during the low cycle of its *STOP* signal. The *READ, WRITE* and $\overline{RESET}$ signals are then generated using logical combinations of different tapped outputs from an inverting delay line that uses the first *STOP* signal as its input. The TDC time-stamps are written to a local memory before being read-out via a shared 16-bit bus on the following excitation cycle.

To allow the timing of multiple photon events per excitation period, an array of multiple TDC pairs is required as well as a method to distribute photon events to this array of timing elements. As described previously, eight timing channels are required per excitation period, meaning that eight TDC pairs, or 2×8 TDCs form the core of the timing architecture.

A pair of interleaved token-passing shift registers, as shown in Fig. 10, are used to distribute events to the array of TDCs.

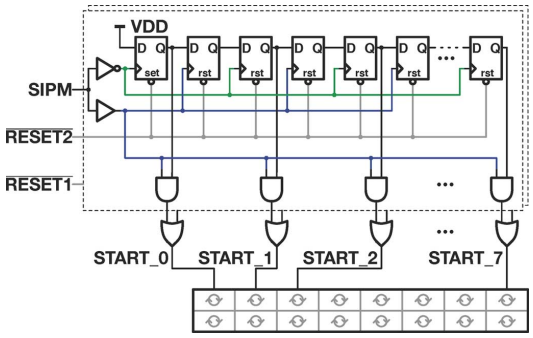


Fig. 10. Token-passing SiPM output event distribution to TDC-pairs.

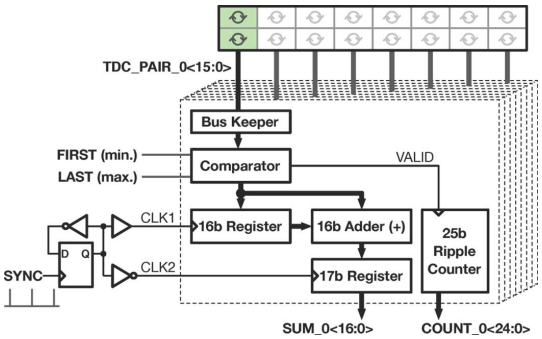


Fig. 11. First-stage CMM calculation.

The token-passing shift register circuits operate in a similar way to the TDC-pairs, where they are active on alternate excitation cycles while the other is reset to prepare it for the next excitation cycle. The shift registers are 15 elements long and alternating bits are clocked using complementary edges of the SiPM output. This ensures that the next TDC-Pair in the array does not have a photon event passed through to it until the previous event has completely finished (signified by its falling edge).

The high-speed asynchronous nature of the photon event arrivals significantly increases the likelihood of flip-flop metastability. To minimize this effect, the token is not cleared from the previous flip-flop in the chain, meaning that if a state goes metastable it will be corrected by the next photon event. Each TDC will only time the first event it sees, so not clearing the token from the previous register does not affect the operation of the architecture.

D. Embedded Processing

Embedding some form of fluorescence lifetime calculation on-chip both miniaturizes the standard experimental set-up and removes the requirement of intense software post-processing of data. Furthermore, by removing the need to send data to a CPU to calculate lifetimes, we introduce the concept of providing lifetime calculations in *real-time*, which opens the way for new applications of time-resolved fluorescence lifetime sensing, such as flow cytometry.

Although *real-time* solutions have previously been performed by using the parallel computing capabilities of FPGAs [27], a large amount of data still needs to be distributed off chip at high rates. The proposed timing architecture creates up to 128-bits (16·8) per excitation cycle, or a data rate of $128 \cdot f$ Mbps (where f is the excitation repetition frequency in MHz). Using an 80 MHz I/O data rate, such throughput would require $1.6 \cdot f$ parallel output pads, or 32 for a conservative 20 MHz excitation rate. As well as not being scalable, such an architecture would require a highly-parallel I/O that would significantly increase I/O power and chip area, creating a severely pad-limited design.

The chosen fluorescence lifetime calculation technique is the centre-of-mass method (CMM) [25], selected for its use of digitized TCSPC data and its high photon efficiency (ideally 100%). As shown by (1), the core of the CMM calculation is the computation of the average TDC code for a given exposure time.

$$\tau_{\mathrm{CMM}} = \left(\frac{\sum_{j=0}^{M-1} j \cdot N_j}{N_C} + \frac{1}{2} \right) \cdot h \qquad (1)$$

where

- $\sum jN_j$ = Sum of valid TDC codes;
- N_C = Count of valid events;
- h = TDC resolution.

Power consumption and area constraints mean that full integer division on-chip is not possible. A power of two photon events could be accumulated, allowing the sum to be binary shifted in order to provide the lifetime estimation calculation. However, performing the division in this way, as implemented on-FPGA for [27], has a major drawback; results are only updated after the power of 2 counts is reached, so the update rate is *photon-rate* dependent and the photon count information is lost. For these reasons, the decision was taken not to perform the division on-chip, but to transfer the total sum and total count to an external device such as an FPGA, micro-controller or even software (depending on the application requirements) to perform the final division, and in doing so preserve the count rate information within the given exposure.

The first stage of the CMM calculation is the summation and counting of events from each pair of TDCs, a block diagram of which is shown in Fig. 11. The data from each set of two TDCs will be driven onto the shared 16-bit bus on alternating synchronization, or clock cycles. In order to provide experimental flexibility, compensate for synchronisation offsets and improve SNR performance, only TDC codes that fall within a pre-defined measurement window are included in the calculation. The window is positioned by calculating CMM on the system's instrument response function (IRF), whilst its width is configured according to the theory developed in [27]. To implement the windowing, a digital comparator is placed on the 16-bit data bus, taking global register values *FIRST* and *LAST*, that define the position and width. Only data that falls inside this window between *FIRST* and *LAST* is passed through to the next stage

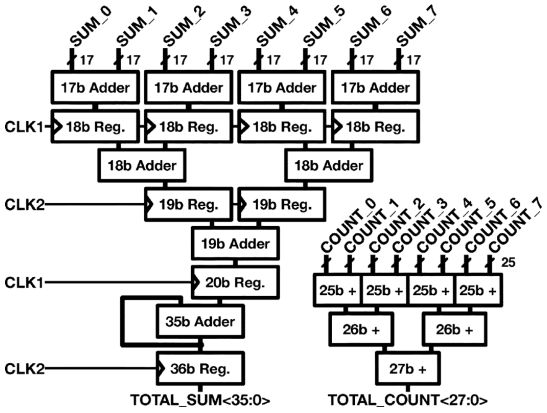


Fig. 12. Final CMM pre-calculation.

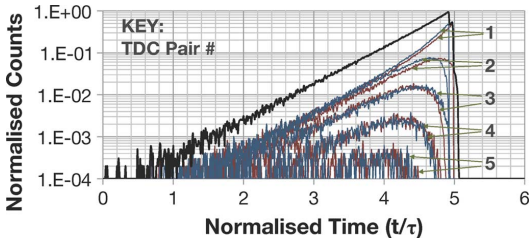


Fig. 13. Modelled TCSPC histogram (black), for 100% *photon-rate* and TDC mismatch of 0.9% showing contribution of each TDC (red and blue).

TABLE I
EXAMPLE TDC RESOLUTIONS (PS) FOR 2 × 8 TDC ARRAY

0	1	2	3	4	5	6	7
50.24	48.98	50.14	49.80	51.61	49.39	50.33	50.32
50.83	50.39	49.41	50.15	51.25	51.37	49.97	49.91

of the calculation. As well as passing data through, the comparator triggers a 25-bit ripple counter to count the number of valid events.

The valid data from each TDC pair is then summed every two excitation cycles. This is implemented by registering the data from one TDC on one clock edge, then storing the result of an addition between it and the data from the second TDC on the inverse clock edge. The *CLK* signals are defined by the same trigger flip-flop as in Fig. 9, as they have the same timing as the TDC *STOP* signals. The sum of each pair of TDCs is then produced on the edge of the second clock signal (*CLK2*).

The final CMM pre-calculation is performed using the data from all 8 TDC pairs, as shown in Fig. 12. Summation of valid time-stamps is carried out using a pipelined adder tree, operating on alternating clock edges, followed by a 36-bit accumulator. The total number of valid TDC events is calculated using a combinatorial adder tree, providing a 28-bit result.

When sampling the CMM data to be read off-chip, the inputs of the TDC architecture need to be gated to allow the pipeline data to settle. Alternated clocking reduces the read-out dead-time by a factor of two. The total sum (36-bit) and count (28-bit) values are then sent off-chip periodically for further integration and used to calculate the final lifetime estimation, corrected for background noise and other calibrated non-idealities such as TDC resolution variation and mismatch.

IV. TIMING NON-IDEALITIES

A. TDC Resolution Variation

The resolution of the TDCs is sensitive to process, voltage and temperature (PVT) variation. It is therefore necessary to be able to calibrate this between devices and over time during an experiment so that lifetime calculations can be corrected accordingly. To facilitate calibration, a ninth TDC pair is added to the device, as shown in Fig. 4, and is used to provide raw TCSPC codes at a maximum rate of one photon event per excitation, but without a processing dead-time.

The timing resolution is then calibrated by configuring a multiplexer in the TDCs to accept an electrical *TEST_START* signal in place of the optical *START* signal from the SiPM. By changing the delay between this signal and the *STOP* signal and capturing histograms, it is possible to measure the TDC resolution at fixed points along the TDC's full range. This calibration measurement is performed periodically and used to correct for environmental changes such as temperature drift and device-to-device mismatch.

Furthermore, the CMM pre-calculation can be configured to process the data from the additional TDC-pair, rather than the multiple-channel timing architecture, to act as a comparison between single and eight-channel timing.

B. TDC Mismatch

Transistor variation also causes mismatch between the resolution of the 16 TDCs in the multiple-channel timing architecture. To study the effect of TDC mismatch on the captured histogram data and CMM lifetime calculation, TDC resolution variation is added to the model introduced in Section III.C. Furthermore, the model is adapted to reflect the reversed start-stop mode used on the device in order to correctly examine the gain-error introduced by the TDC-mismatch which causes variation of the histogram peak between TDCs.

A 0.9% (≈ 0.45 ps) standard-deviation of the resolution mismatch is quoted for a 512 array of the same TDC [26]. The histograms in Fig. 13 show an example of the contribution that each TDC in the 2×8 timing architecture (1–5) makes to the final histogram for a *photon-rate* of 100% of the excitation rate. In the example, a 50 ps mean, 0.45 ps standard deviation TDC resolution is used to perform the simulation, as shown in Table I. The mismatch causes a systematic error in the CMM lifetime calculation, in this case $\approx 6\%$, that is dependent on the recorded CMM data (sum and counts) and can be calibrated using known fluorescent samples. A look-up table (LUT) is then employed to correct for the error introduced.

V. EXPERIMENTAL SETUP

The device is evaluated using an Opal Kelly XEM-3010 FPGA and USB platform mounted on a custom printed circuit

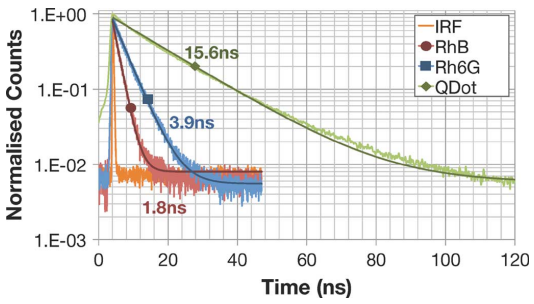


Fig. 14. Raw TCSPC histogram results with extracted lifetimes.

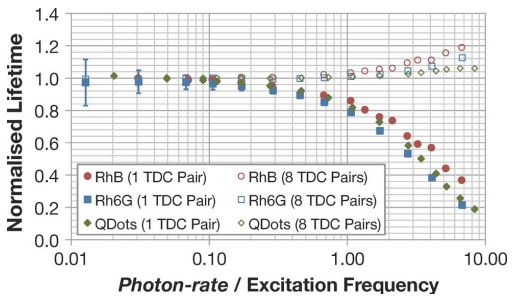


Fig. 15. CMM results showing resistance to *pile-up*.

board. For experimental evaluation of the sensor's performance, a range of fluorescent dyes are measured using varying laser excitation power. Although the sensor acts as a point detector, for experimental convenience measurements are performed as part of an existing wide-field fluorescence lifetime imaging system on a Nikon TE2000U inverted microscope. The excitation source is a PicoQuant pulsed diode laser with a wavelength of 478 nm coupled through the epi-fluorescence port of the microscope using a Nikon B-2A filter cube. The laser pulse repetition rate is 10 MHz and the maximum power reaching the back focal plane of the objective is $\approx 64\ \mu$W. The *photon-rate* can be varied by combining various neutral density filters to adjust the intensity of the excitation laser.

The excited sample volume is focussed onto the active area of the CMOS device, which is attached to one of the camera ports using an additional short focal length lens. Light was concentrated on a set of eight SPADs, arranged within a 3×3 grid, so that each individual SPAD obtains a similar count rate. The SPADs were biased with $V_{\mathrm{EB}} = 3.0$ V and $V_{\mathrm{BD}} = -13.4$ V, providing a maximum count rate in excess of 100 MHz and a combined DCR of 5.5 kHz. Before acquiring each set of experimental data, the TDC resolution is calibrated, giving an average value of 56.5 ps throughout experiments.

VI. Results

To fully evaluate the system, the device is tested in a number of the following configurations:

- Single-channel timing for TCSPC histogramming;
- Single-channel timing with embedded CMM calculation;
- Multi-channel timing with embedded CMM calculation.

Raw TCSPC histograms were initially acquired from 1 second exposures captured using the validation TDC pair outputting raw time-stamps for different fluorophores. Aqueous solutions of Rhodamine B (1 mM) and Rhodamine 6G (1 mM) as well as Birch Yellow Quantum Dot in toluene (60 μM - Evident Technologies, NY, USA), which have quoted lifetimes of 1.74 ns, 4.08 ns and 15–20 ns respectively, were used. The captured lifetime decay histograms are shown together with the IRF and fitted curves in Fig. 14. The FWHM of the IRF is measured as 325 ps at a mean TDC code of 25 ns. Lifetimes of 1.8 ns, 3.9 ns and 15.6 ns were calculated for the fluorophores with Edinburgh Instruments FAST software. This result highlights the ability of the device to operate as a typical TCSPC acquisition system, with the added benefit of having no processing dead-time. However, the remaining *pile-up* limitations of standard TCSPC implementations still exist.

To further miniaturize and speed-up the typical fluorescence lifetime setup, the embedded CMM pre-calculation is enabled using the data produced by the same single timing element (validation TDC-pair) as above. The CMM calculation, which removes the requirement of intense CPU processing of histogram data, successfully produces lifetime estimates of 1.7 ns, 3.9 ns and 16.5 ns after background correction [25], which are in good agreement with the TCSPC results for the same fluorophores.

Further background corrected CMM calculations were then performed on the same fluorophores with an increasing *photon-rate* for both single and eight-channel timing elements available per excitation period. The normalized CMM lifetime results are shown in Fig. 15 for all fluorophores, clearly showing the device's ability to significantly overcome the *pile-up* limitation for count rates in excess of the excitation rate, or over 10 MHz in this case. This represents more than an order of magnitude improvement over typical TCSPC systems, that would only be capable of acquiring count rates of up to 1 MHz [4], without the requirement for pile-up correction. Even assuming an equivalent sub-500 ps discrete detector output pulse-width, an expensive, complex and power hungry system of more than 10 discrete timing channels would be necessary to achieve a comparable throughput. The results in Fig. 15 follow the modelled data very closely, both for single and eight timing-channels, as can be seen by comparing Rhodamine 6G with the 4 ns lifetime simulation results in Fig. 7. Furthermore, the effect of the SiPM pulse-width as a ratio of lifetime is highlighted, with shorter lifetime fluorophores (Rhodamines) diverging from their lifetime faster than the longer lifetime fluorophore (Quantum dots) at high *photon-rates*.

Finally, an experiment was run using multi-channel timing and embedded CMM calculation using a mixture of two types of fluorescently labeled polystyrene beads. Commercial yellow-green fluorescent 2.4 μm diameter beads (Interfacial Dynamics Corporation, Portland, OR) and 2 μm diameter beads labeled with fluorescein have distinct lifetimes ($\approx$ 7ns and 2.7 ns, resp.). The mixed sample was scanned through the laser focus whilst 4ms exposures were acquired. The results, as shown in Fig. 16, clearly show the device's ability to distinguish these two labeled beads. The brighter sample produced a count rate of ≈ 2.5 MHz,

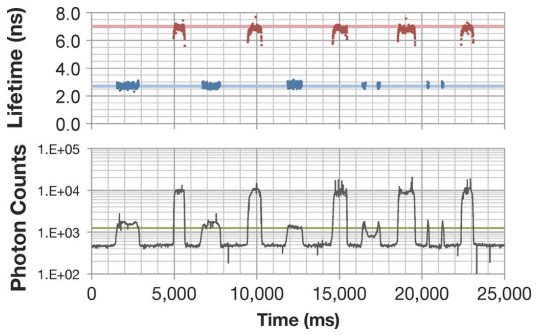


Fig. 16. CMM results showing distinction between two fluorophores.

a *photon-rate* of 25% of the excitation rate, further highlighting the device's resistance to *pile-up*. If this experiment was run using a typical TCSPC setup, laser power would be tuned to provide a *photon-rate* of $\approx$ 5% of the excitation rate for the brightest sample, significantly reducing the photon throughput of both samples by a factor of 5.

Device core power consumption is *photon-rate* dependent, ranging from $\approx$ 1.8 mW in the dark, up to $\approx$ 9.5 mW at the highest intensities used in the experiments presented.

VII. Conclusion

A miniaturized smart fluorescence lifetime sensor, capable of processing photon count rates well beyond the classic TCSPC *pile-up* limit, without the need of *pile-up* correction, has been presented and demonstrated experimentally. The ability of the device to operate over a much higher dynamic range, both in terms of *photon-rate* and lifetime, paves the way for improvements in two-photon or confocal scanning FLIM. Furthermore, by combining the device's high photon-throughput with its ability to perform fluorescence lifetime estimation calculations in *real-time* and with low latency, time-resolved fluorescence lifetime based flow cytometry or sorting [28] for high throughput screening, is made feasible.

Acknowledgment

The authors would like to thank STMicroelectronics for chip fabrication. The assistance of R. Walker is gratefully acknowledged.

References

[1] A. Esposito, "Beyond range: Innovating fluorescence microscopy," *Remote Sens.*, vol. 4, no. 1, pp. 111–119, Jan. 2012.

[2] R. Sanders, A. Draaijer, H. C. Gerritsen, P. M. Houpt, and Y. K. Levine, "Quantitative pH imaging in cells using confocal fluorescence lifetime imaging microscopy," *Anal. Biochem.*, vol. 227, pp. 302–308, Sep. 1995.

[3] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, and M. L. Johnson, "Fluorescence lifetime imaging of calcium using quin-2," *Cell Calcium*, vol. 13, no. 3, pp. 131–147, Mar. 1992.

[4] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*. New York: Springer, 2005.

[5] D. V. O. Connor and D. Phillips, *Time-correlated Single Photon Counting*. New York: Academic, 1984.

[6] M. Köllner and J. Wolfrum, "How many photons are necessary for fluorescence-lifetime measurements?," *Chem. Phys. Lett.*, vol. 200, no. 1–2, pp. 199–204, Nov. 1992.

[7] J. Yguerabide, "Nanosecond fluorescence spectroscopy of macromolecules," *Methods Enzymol.*, vol. 26, pp. 498–578, 1972.

[8] C. M. Harris and B. K. Selinger, "Single-photon decay spectroscopy II—The pile-up problem," *Aust. J. Chem.*, vol. 32, no. 10, pp. 2111–2129, 1979.

[9] D. McLoskey, D. Campbell, A. Allison, and G. Hungerford, "Fast time- correlated single-photon counting fluorescence lifetime acquisition using a 100 MHz semiconductor excitation source," *Meas. Sci, Technol.*, vol. 22, no. 6, p. 067001, Apr. 2011.

[10] J. R. Lakowicz, *Principals of Fluorescence Spectroscopy*, 3rd ed. New York: Springer, 2006.

[11] J. P. Houston, M. A. Naivar, and J. P. Freyer, "Digital analysis and sorting of fluorescence lifetime by flow cytometry," *Cytometry A*, vol. 77, no. 9, pp. 861–872, Sept. 2010.

[12] E. Gratton, S. Breusegem, J. Sutin, Q. Ruan, and N. Barry, "Fluorescence lifetime imaging for the two-photon microscope: Time-domain and frequency-domain methods," *J. Biomed. Opt.*, vol. 8, no. 3, pp. 381–90, Jul. 2003.

[13] R. P. Hertzberg and A. J. Pope, "High-throughput screening: new technology for the 21st century," *Curr. Opin. Chem. Biol.*, vol. 4, no. 4, pp. 445–451, Aug. 2000.

[14] D. Contini, A. Torricelli, A. Pifferi, L. Spinelli, F. Paglia, and R. Cubeddu, "Multi-channel time-resolved system for functional near infrared spectroscopy," *Opt. Exp.*, vol. 14, no. 12, pp. 5418–5432, Jun. 2006.

[15] P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kantzerov, V. Kaplin, A. Karakash, F. Kayumov, S. Klemin, E. Popova, and S. Smirnov, "Silicon photomultiplier and its possible applications," *Nucl. Instrum. Methods Phys. Res.. A*, vol. 504, no. 1–3, pp. 48–52, May 2003.

[16] C. Veerappan, J. Richardson, R. Walker, D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160128 single-photon image sensor with on- pixel 55 ps 10b time-to-digital converter," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 2011, pp. 312–313.

[17] D. Mosconi, D. Stoppa, L. Pancheri, L. Gonzo, and A. Simoni, "CMOS single-photon avalanche diode array for time-resolved fluorescence detection," in *Proc. Eur. Solid State Circuits Conf.*, Sep. 2006, pp. 564–567.

[18] B. R. Rae, J. Yang, J. McKendry, Z. Gong, D. Renshaw, J. M. Girkin, E. Gu, M. D. Dawson, and R. K. Henderson, "A vertically integrated CMOS microsystem for time-resolved fluorescence analysis," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 6, pp. 437–444, Dec. 2010.

[19] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable single-photon avalanche diode structures in nanometer CMOS technology," *IEEE Trans. Electron Devices*, vol. 58, no. 7, pp. 2028–2035, Jul. 2011.

[20] L. H. C. Braga, L. Pancheri, L. Gasparini, M. Perenzoni, R. Walker, R. K. Henderson, and D. Stoppa, "A CMOS mini-SiPM detector with in-pixel data compression for PET applications," in *Proc. Nuclear Science Symp. Conf. Rec.*, Oct. 2011, pp. 548–552.

[21] P. B. Coates, "Pile-up corrections in the measurement of lifetimes," *J. Phys. E*, vol. 5, no. 2, pp. 148–150, Feb. 1972.

[22] O. M. Williams and W. J. Sandle, "A pile-up gate generator for removing distortion in multichannel delayed coincidence experiments," *J. Phys. E, Sci. Instrum.*, vol. 3, no. 9, pp. 741–743, Sep. 1970.

[23] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, "A 100 Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13 m CMOS imaging technology," in *Proc. Int. Solid-State Circuits Conf.*, Feb. 2012, pp. 122–123.

[24] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier principle of operation and intrinsic detector performance," in *Proc. IEEE Nuclear Science Symp. Conf. Rec.*, Oct. 2009, pp. 1959–1965.

[25] D.-U. Li, B. Rae, R. Andrews, J. Arlt, and R. Henderson, "Hardware implementation algorithm and error analysis of high-speed fluorescence lifetime sensing systems using center-of-mass method," *J. Biomed. Opt.*, vol. 15, no. 1, p. 017006, Jan. 2010.

[26] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, "A 3232 50 ps resolution 10 bit time to digital converter array in 130 nm CMOS for time correlated imaging," in *Proc. Custom Integrated Circuits Conf.*, Sep. 2009, no. 029217, pp. 77–80.

[27] D. D.-U. Li, J. Arlt, D. Tyndall, R. Walker, J. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm," *J. Biomed. Opt.*, vol. 16, no. 9, p. 096012, Sep. 2011.

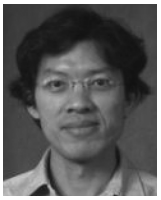
[28] C. D. Salthouse, R. Weissleder, and U. Mahmood, "Development of a time domain fluorimeter for fluorescent lifetime multiplexing analysis," *IEEE Trans. Biomed. Circuits Syst.*, vol. 2, no. 3, pp. 204–211, Sep. 2008.

David Tyndall (S'12) received the M.Eng. (Joint Honors) degree in electronics and software engineering from The University of Edinburgh, Edinburgh, U.K., in 2009.

He is currently working toward the Ph.D. degree at The University of Edinburgh, funded by the U.K. Engineering and Physical Sciences Research Council. His interests include researching high throughput time correlated imaging and sensing for bio applications. Since August 2012, he has been with Dialog Semiconductor, Edinburgh, U.K. He was awarded the International Engineering Consortium William L. Everitt Prize for his studies.

Bruce R. Rae (M'08) received the M.Eng. and Ph.D. degrees in electrical and electronic engineering from The University of Edinburgh, Edinburgh, U.K., in 2005 and 2009, respectively.

During work on his Master's degree, he worked at STMicroelectronics Imaging Division, where he has been an Analogue Design Engineer since March 2011. His doctoral project focused on the design and implementation of a low-cost, miniaturized CMOS-based micro-system for time-resolved fluorescence analysis. From September 2008 until March 2011, he was a postdoctoral Research Associate at the Institute for Integrated Micro and Nano Systems at The University of Edinburgh, School of Engineering, with research interests in CMOS single photon counting and micro-LED devices.



David Day-Uei Li received the Ph.D. degree in optical waveguide devices from National Taiwan University, Taipei City, Taiwan, in 2001.

He joined the SoC Technology Center, Industrial Technology Research Institute, Taiwan, as an R&D Engineer, working on transceivers for 10 Gigabit/s fiber-optical communication systems. In March 2007, he joined the Institute for Integrated Micro and Nano Systems, University of Edinburgh, working on the European projects for single-photon avalanche diode (SPAD) based fluorescence lifetime imaging microscopy (FLIM) cameras and analogue front-end circuits for oxygen/pH sensing. His work resulted in the first video-rate FLIM imaging on CMOS SPAD arrays. In May 2011, he joined the School of Engineering and Design, University of Sussex as a Lecturer in Biomedical Engineering. His current research interests include mixed-signal integrated circuits design, CMOS SPAD-based FLIM cameras, hardware/algorithms for fluorescence based sensing and imaging, finite element and finite difference analysis, instrumentation for breast cancer diagnosis, numerical methods and optical fiber/waveguide components and electronic circuits modeling.

Jochen Arlt received the Ph.D. degree from the University of St. Andrews, Scotland, U.K., in 2001.

He is Lab Manager and Research Fellow of the Collaborative Optical Spectroscopy, Micromanipulation and Imaging Centre (COSMIC) at The University of Edinburgh. He has extensive expertise in optical imaging techniques and novel instrumentation, with particular focus on optical tweezers and fluorescence techniques, which he applies to a broad range of interdisciplinary projects in soft-condensed matter and biophysics.

Abigail Johnston is an undergraduate M.Phys. student at The University of Edinburgh, Edinburgh, U.K.

She completed a Senior Honors Project entitled *Fast Fluorescence Lifetime Detection*, under the supervision of Dr. Jochen Arlt. Following the project, she continued her Master's studies and traveled to Germany to undertake a summer placement at the Max-Planck-Institute for Dynamics and Self-Organization, Göttingen.

Justin A. Richardson (M'01) was born in Ann Arbor, MI, in 1970. He received the Ph.D. degree from The University of Edinburgh, Edinburgh, U.K., in 2010.

His doctoral dissertation was entitled, *Time Resolved Single Photon Imaging in Nanometer Scale CMOS Technology*. During his Ph.D. research, he specialized in the design and development of single-photon avalanche diodes and time-to-digital converter arrays for the European *Megaframe* project.. He was a Research Associate with The University of Edinburgh and maintained employment with STMicroelectronics Imaging Division. Since January 2011, he has been with Dialog Semiconductor, Edinburgh, and has been a Visiting Researcher with The University of Edinburgh. He is the author of more than 25 publications and is the holder of nine patents. His research interests include single-photon detectors, as well as analog- and time-to-digital conversion techniques.

Robert K. Henderson (M'84) received the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 1990.

Currently, he is a Reader at the School of Engineering in the Institute for Microelectronics and Nanosystems, The University of Edinburgh, Edinburgh, U.K. Since 1991, he has been a Research Engineer at the Swiss Centre for Microelectronics, Neuchatel, Switzerland, working on low-power sigma-delta analog-to-digital converters and digital-to-analog converters for portable electronic systems. In 1996, he was appointed Senior VLSI Engineer at VLSI Vision Ltd., Edinburgh, where he worked on the world's first single-chip video camera and was Project Leader for many other complementary metal-oxide semiconductor (CMOS) image sensors. Since 2000, as Principal VLSI Engineer in the ST Microelectronics Imaging Division, he led the design of the first image sensors for mobile phones, resulting in annual revenues of several hundred million dollars. He joined the University of Edinburgh in 2005 to pursue his research interests in CMOS integrated-circuit design, imaging, and biosensors. As PI on the joint European project MegaFrame with three European Universities and ST Microelectronics, he has led research resulting in the first single-photon avalanche diode in nanometer CMOS technology. He is the author of 104 papers and 17 patents.

Dr. Henderson was awarded Best Paper Award at the 1996 European Solid-State Circuits Conference as well as the 1990 IEE J. J. Thomson Premium.

B.3 Tyndall et. al., ISSCC, 2012

ISSCC 2012 / SESSION 6 / MEDICAL, DISPLAYS AND IMAGERS / 6.7

6.7 A 100Mphoton/s Time-Resolved Mini-Silicon Photomultiplier with On-Chip Fluorescence Lifetime Estimation in 0.13µm CMOS Imaging Technology

David Tyndall[1], Bruce Rae[2], David Li[3], Justin Richardson[4], Jochen Arlt[1], Robert Henderson[1]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]STMicroelectronics, Edinburgh, United Kingdom
[3]University of Sussex, Brighton, United Kingdom
[4]Dialog Semiconductor, Edinburgh, United Kingdom

Time-correlated single photon counting (TCSPC) is a technique whereby low-light signals are recorded with picosecond timing resolution relative to a synchronized optical impulse excitation, in order to extract the characteristic fluorescence decay constant, or lifetime [1]. Typical TCSPC apparatus includes a pulsed optical source, a discrete detector such as an avalanche photodiode (APD) or photomultiplier tube (PMT), external time-to-digital conversion (TDC) hardware and a PC to compute the decay constant, resulting in a bulky, expensive and power-hungry acquisition system. A major limitation of this approach is the restrictively low photon count limit of 1-to-5% of the excitation rate, which is necessary in order to avoid distortion due to photon 'pile-up' caused by both long detector dead-time and the inability of the TDC hardware to process more than one event per excitation period. As such, promising applications of TCSPC including cell cytometry, confocal microscopy, high throughput screening (HTS), and functional near infrared spectroscopy (fNIRS) are severely limited by peak acquisition rates of 1MHz. Although 100MHz has been achieved [2], the approach used is restricted to fluorescent dyes with lifetimes less than 2ns. Recent advances in single-photon avalanche diodes (SPADs) and on-chip TDCs manufactured in standard CMOS processes have enabled TCSPC measurements to be performed by an imaging array [3]; however such devices produce data at over 25Gb/s, have low fill factors of ~2% and pixel update rates are limited. Time-gated lifetime sensing significantly reduces the data bandwidth and processing time [4,5], but is photon inefficient and still limited by pile-up.

In this paper, we present a fully digital single-chip solution to fluorescence lifetime sensing implementing lifetime estimation on-chip using the centre-of-mass method (CMM) [6]. A time-multiplexed, multichannel TDC architecture is introduced to allow the pile-up limit to be broken, achieving a maximum 100Mphoton/s acquisition rate, whilst still allowing lifetime decays of common organic fluorophores to be obtained. The device comprises a 1.3×1.7mm^2 CMOS chip in 0.13µm technology, featuring an array of SPAD detectors in a silicon photomultiplier (SiPM) architecture, the multichannel TDC architecture, a CMM processing block and a serial interface for control and data capture (Fig. 6.7.1).

The mini-SiPM array comprises 32×32 8µm diameter active area passively quenched SPADs from [3] with a 22.5µm pitch, providing a total active area of over 0.05mm^2 with a 10% fill factor. To prevent the dead-time from restricting the maximum count rate, a pulse-shortening monostable circuit is included at the output of each SPAD to reduce the pulse width to approximately 2ns (Fig. 6.7.2) [7]. This increases the possible photon throughput by an order of magnitude, with a maximum recorded count rate of over 350MHz. These shorter pulses then pass through a balanced OR-tree to provide a single output to the time-resolving circuitry. Additionally, individual SPADs can be disabled independently, allowing high dark count detectors to be switched off. Furthermore, a separate high voltage SPAD supply is used that can be adjusted independently of the device core voltage, protecting the core circuitry from light dependent power consumption. Both of these mechanisms are also used to control the sensitivity of the SiPM.

Time-resolved measurements are performed using TDCs with a ~52ps resolution, modified from [3] to allow an extended range of up to 16 bits (~3.6µs), enabling the resolvability of longer lifetime fluorophores. The TDCs are arranged in pairs with interleaved timing, allowing one TDC to remain active and available to accept a photon arrival while the other is writing its data and being reset (Fig. 6.7.3). The excitation synchronization pulse defines the timing for these TDC pairs so that a single resolved time-stamp is generated per-pair for every excitation cycle. To achieve this, a T-type flip-flop is used to derive a pair of anti-phase clocks from the synchronization signal, from which additional timing signals are defined. The TDC time-stamps are written to a local memory before being read-out via a shared 16b bus on the following excitation cycle, where they are summed as the first stage of the CMM calculation.

To allow generation of multiple time-stamps per excitation, an array of 8 TDC pairs is used. A token-ring distributes each SPAD pulse to an idle TDC (Fig. 6.7.4). This consists of a 16b shift register, where alternate bits are triggered using complementary edges of SPAD events. It is therefore possible to capture up to 8 photons per excitation and as such the 2×8 TDC architecture is capable of processing up to 100Mphoton/s with an excitation repetition rate of 12.5MHz.

On-chip lifetime estimation is performed using CMM, which calculates the center of mass of the decay histogram by averaging TDC time-stamps. Addition of time-stamps is performed using a pipelined adder tree followed by an accumulator, whilst the number of TDC events is counted using ripple counters local to the TDC pairs, which are summed using a combinatorial adder tree. The photon events are registered and their time-stamps are summed if they fall within a measurement window defined by the control registers FIRST and LAST. This windowing is implemented using a digital comparator at the output of each TDC pair (Fig. 6.7.3). The total sum (36b) and count (28b) values are then sent off-chip periodically for further integration and used to calculate the final lifetime estimation, corrected for background noise.

For data validation purposes, one TDC pair is configured to output raw TDC time-stamps, which are captured and used to build a decay histogram in software. This mode is also used as a calibration technique by configuring the TDCs to accept a test START signal, which has a known and fixed offset from the STOP signal. This calibration measurement is performed periodically and used in software to correct for environmental changes such as temperature drift and device-to-device mismatch.

Both raw histograms and CMM estimations were acquired for bulk samples of Rhodamine B (1mM), Rhodamine 6G (1mM) and Birch Yellow Quantum Dot (60µM) fluorophores, which have quoted lifetimes of 1.68ns, 4.08ns and 15-20ns respectively. Before acquiring each set of experimental data, a TDC resolution of 52.4ps was measured using the calibration technique described above. Lifetimes of 1.7ns, 3.9ns and 15.5ns where calculated for the fluorophores with Edinburgh Instruments FAST software from 1 second exposure histograms captured using the validation TDC pair outputting raw time-stamps (Fig. 6.7.5). Background corrected CMM calculations were performed with multiple 1ms exposures, producing lifetime estimations of 1.7ns, 3.9ns and 17.9ns (Fig. 6.7.6). These values are in good agreement with the quoted lifetimes for the fluorophores used.

Acknowledgements:
This work was supported by the UK EPSRC. The authors would like to thank STMicroelectronics for chip fabrication. The assistance of Richard Walker is gratefully acknowledged.

References:
[1] W. Becker, "Advanced Time-Correlated Single Photon Counting Techniques," *Springer Series in Chemical Physics,* Vol. 81, 2005.
[2] D. McLoskey, *et al.*, "Fast Time-Correlated Single-Photon Counting Fluorescence Lifetime Acquisition using a 100 MHz Semiconductor Excitation Source," *Meas. Sci. Technol. 22, 067001,* 2011.
[3] C. Veerappan, *et al.*, "A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter," *ISSCC Dig. Tech. Papers,* pp. 312-313, Feb 2011.
[4] B.R. Rae, *et al.*, "A Microsystem for Time-Resolved Fluorescence Analysis using CMOS Single-photon Avalanche Diodes and Micro-LEDs," *ISSCC Dig. Tech. Papers,* pp. 166-167, Feb. 2008.
[5] D. Mosconi, *et al.*, "CMOS Single Photon Avalanche Diode Array for Time-Resolved Fluorescence Detection," *IEEE Proc. of ESSCIRC,* pp. 564-567, Sept. 2006.
[6] D.-U. Li, *et al.*, "Hardware Implementation Algorithm and Error Analysis of High-Speed Fluorescence Lifetime Sensing Systems using Center-of-Mass Method," *J. Biomed. Opt. 15, 017006,* Feb. 2010.
[7] L.H.C. Braga, *et al.*, "A CMOS Mini-SiPM Detector with in-Pixel Data Compression for PET Applications," *IEEE Nuclear Science Symp.,* 2011.

 978-1-4673-0377-4/12/$31.00

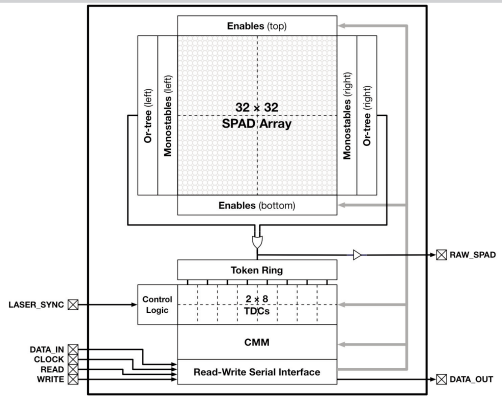


Figure 6.7.1: System overview block diagram, showing SiPM Array with enables and pulse shortening monostable circuits, time-resolving TDCs, CMM algorithm and serial interface.

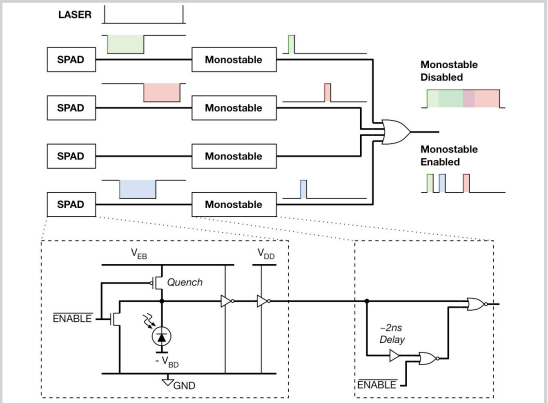


Figure 6.7.2: SPAD with independent supply (VEB), pulse shortening monostable circuit and output OR-tree showing discrimination of multiple overlapping events [7].

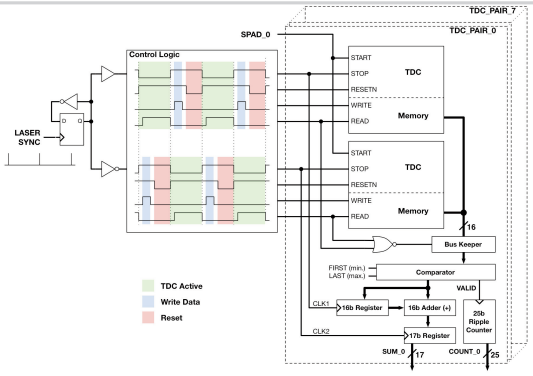


Figure 6.7.3: Timing generation for interleaved TDC pairs, created from the excitation synchronization pulse. The TDCs operate with reverse mode timing [1], where the SPAD event starts the TDC and the synchronization pulse stops the TDC.

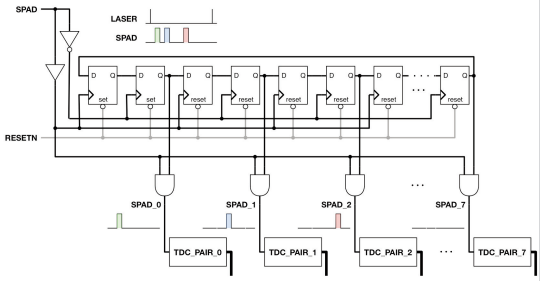


Figure 6.7.4: Token-ring for SPAD pulse distribution to TDC pairs, showing three consecutive events being sent to idle TDC pairs.

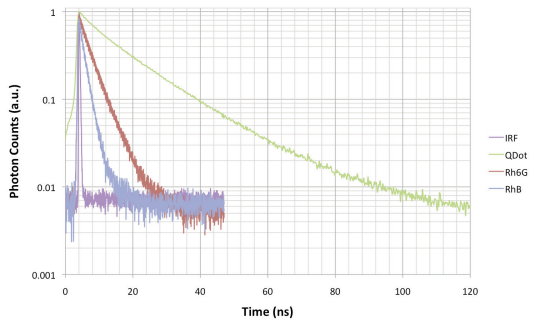


Figure 6.7.5: Normalized histograms of Rhodamine B, Rhodamine 6G, Birch Yellow Quantum Dot and instrument response function (IRF) captured using the on-chip verification TDCs. Lifetime values of 1.7ns, 3.9ns and 15.5ns were calculated in software for these decays.

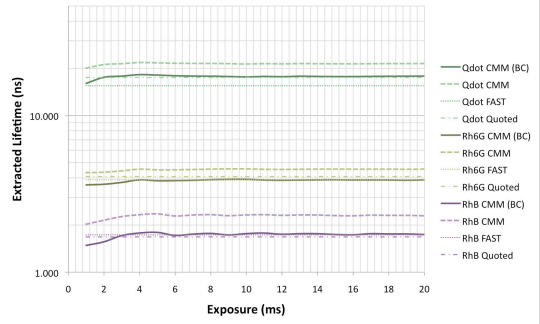


Figure 6.7.6: Comparison of on-chip computed lifetime estimations by background corrected (BC) and uncorrected CMM compared with quoted and software extracted values form Fig. 6.7.5.

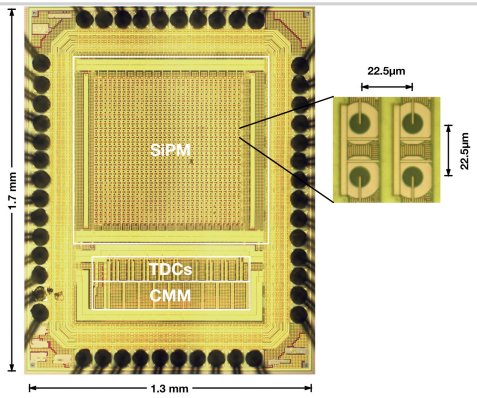


Figure 6.7.7: Micrograph of chip in 48CLCC package, measuring 1.3×1.7mm and showing the SiPM, TDC and CMM blocks. Inset: SPAD pitch of 22.5µm providing a fill factor of over 10%.

978-1-4673-0377-4/12/$31.00